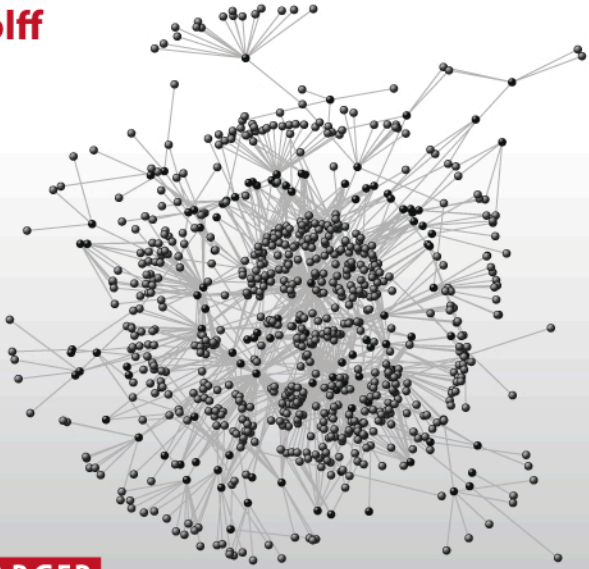
Genome Dynamics
Editor: J.-N. Volff
Vol. 3

Gene and Protein Evolution

Editor
J.-N. Volff



KARGER

VISIT…

Genome Dynamics

Vol. 3

Series Editor

Jean-Nicolas Volff, Lyon

Executive Editor

Michael Schmid, Würzburg

Advisory Board

John F.Y. Brookfield, Nottingham
Jürgen Brosius, Münster
Pierre Capy, Gif-sur-Yvette
Brian Charlesworth, Edinburgh
Bernard Decaris, Vandoeuvre-lès-Nancy
Evan Eichler, Seattle, WA
John McDonald, Atlanta, GA
Axel Meyer, Konstanz
Manfred Schartl, Würzburg

Gene and Protein Evolution

Volume Editor

Jean-Nicolas Volff, Lyon

34 figures, 18 in color, and 10 tables, 2007

KARGER

Basel · Freiburg · Paris · London · New York ·
Bangalore · Bangkok · Singapore · Tokyo · Sydney

Prof. Jean-Nicolas Volff
Institut de Génomique Fonctionnelle de Lyon
Ecole Normale Supérieure de Lyon
46 allée d'Italie
F-69364 Lyon Cedex 07 (France)

Library of Congress Cataloging-in-Publication Data

Gene and protein evolution / volume editor, Jean-Nicolas Volff.
p. ; cm. – (Genome dynamics, ISSN 1660-9263 ; v. 3)
Includes bibliographical references and indexes.
ISBN-13: 978-3-8055-8340-4 (hard cover : alk. paper)
1. Genomics. 2. Molecular evolution. 3. Proteins–Evolution. I. Volff, Jean-Nicolas. II. Series.
[DNLM: 1. Genomics. 2. Evolution, Molecular. 3. Proteins. QU 58.5 G3255 2007]
QH447.G44 2007
572.8′38–dc22

2007024911

Bibliographic Indices. This publication is listed in bibliographic services, including Current Contents® and Index Medicus.

Disclaimer. The statements, options and data contained in this publication are solely those of the individual authors and contributors and not of the publisher and the editor(s). The appearance of advertisements in the book is not a warranty, endorsement, or approval of the products or services advertised or of their effectiveness, quality or safety. The publisher and the editor(s) disclaim responsibility for any injury to persons or property resulting from any ideas, methods, instructions or products referred to in the content or advertisements.

Drug Dosage. The authors and the publisher have exerted every effort to ensure that drug selection and dosage set forth in this text are in accord with current recommendations and practice at the time of publication. However, in view of ongoing research, changes in government regulations, and the constant flow of information relating to drug therapy and drug reactions, the reader is urged to check the package insert for each drug for any change in indications and dosage and for added warnings and precautions. This is particularly important when the recommended agent is a new and/or infrequently employed drug.

All rights reserved. No part of this publication may be translated into other languages, reproduced or utilized in any form or by any means electronic or mechanical, including photocopying, recording, microcopying, or by any information storage and retrieval system, without permission in writing from the publisher.

© Copyright 2007 by S. Karger AG, P.O. Box, CH–4009 Basel (Switzerland)
www.karger.com
Printed in Switzerland on acid-free and non-aging paper (ISO 9706) by Reinhardt Druck, Basel
ISSN 1660–9263
ISBN 978–3–8055–8340–4

Contents

VII Preface

1 Coevolution within and between Genes
Galtier, N.; Dutheil, J. (Montpellier)

13 Evolution of Protein-Protein Interaction Network
Makino, T. (Mishima/Shizuoka/Dublin); Gojobori, T. (Mishima/Tokyo)

30 Bacterial Flagella and Type III Secretion: Case Studies in the Evolution of Complexity
Pallen, M.J. (Birmingham); Gophna, U. (Tel-Aviv)

48 Comparative Genomics and Evolutionary Trajectories of Viral ATP Dependent DNA-Packaging Systems
Burroughs, A.M. (Bethesda, Md./Boston, Mass.); Iyer, L.M.; Aravind, L. (Bethesda, Md.)

66 General Trends in the Evolution of Prokaryotic Transcriptional Regulatory Networks
Madan Babu, M. (Bethesda, Md./Cambridge); Balaji, S.; Aravind, L. (Bethesda, Md.)

81 Divergence of Regulatory Sequences in Duplicated Fish Genes
Van Hellemont, R. (Leuven); Blomme, T.; Van de Peer, Y. (Ghent); Marchal, K. (Leuven)

101 Evolution of Gene Function on the X Chromosome Versus the Autosomes
Singh, N.D.; Petrov, D.A. (Stanford, Calif.)

119 Amino Acid Repeats and the Structure and Evolution of Proteins
Albà, M.M. (Barcelona); Tompa, P. (Budapest); Veitia, R.A. (Paris)

131 Origination of Chimeric Genes through DNA-Level Recombination
Arguello, J.R.; Fan, C. (Chicago, Ill.); Wang, W. (Kunming); Long, M. (Chicago, Ill.)

147 Exaptation of Protein Coding Sequences from Transposable Elements
Bowen, N.J.; Jordan, I.K. (Atlanta, Ga.)

163 Modulation of Host Genes by Mammalian Transposable Elements
Makałowski, W. (University Park, Ill.); Toda, Y. (Tokyo)

175 Modern Genomes with Retro-Look: Retrotransposed Elements, Retroposition and the Origin of New Genes
Volff, J.-N. (Lyon); Brosius, J. (Münster)

191 Author Index

192 Subject Index

Preface

The third volume of "Genome Dynamics" is dedicated to "Gene and Protein Evolution". Relatively recently, the genomics era has completely changed our way to apprehend evolution, particularly through the emergence of comparative genomics, a discipline allowing the analysis of complete genomes and biological processes over huge periods of time. In this volume, a panel of internationally recognized experts present and discuss an update of the evolutionary processes at the basis of organismal diversification and complexity, and review the mechanisms leading to the acquisition of new traits and new functions. Different levels of evolution will be considered, from internal modules in genes and proteins to interactomes and biological networks, with integration of the influence of both the genomic environment and the ecological context. Particular emphasis will be given to the origin of novel genes and gene functions, as well as to the evolutionary impact of the duplication of genetic information, with several chapters devoted to transposable elements.

All papers published in Genome Dynamics are reviewed according to classical standards. I would like to thank all contributors and referees involved in this book, Michael Schmid and his team, as well as Karger Publishers for their invaluable help during the preparation of this volume.

Jean-Nicolas Volff
Lyon, June 2007

Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 1–12

Coevolution within and between Genes

N. Galtier, J. Dutheil

CNRS UMR 5171 – Génome, Populations, Interactions, Adaptation, Université Montpellier 2, France

Abstract

Interacting biological systems do not evolve independently, as exemplified many times at the cellular, organismal and ecosystem levels. Biological molecules interact tightly, and should therefore coevolve as well. Here we review the literature about molecular coevolution, between residues within RNAs or proteins, and between proteins. A panel of methodological and bioinformatic approaches have been developed to address this issue, yielding contrasting results: a strong coevolutionary signal is detected in RNA stems, whereas proteins show only moderate, uneasy to interpret departure from the independence hypothesis. The reasons for this discrepancy are discussed.

Copyright © 2007 S. Karger AG, Basel

Two or more biological systems are considered to be coevolving when they do not evolve independently from each other, i.e. when changes occurring in system 1 influence system 2, modifying the probabilities of the future states it might take. Coevolution obviously occurs between interacting species, such as hosts and pathogens, or symbionts. Immune systems undergo diversifying evolution as a response to viral and bacterial invention of new attacking systems, an arm race process recalling Lewis Carroll's Red Queen [1]. Within species, the male and female reproductive apparatus or mating behavior, for instance, coevolve tightly [2, 3]. At the cellular level, coadaptation has been demonstrated between the nuclear and mitochondrial genomes: cybrids, i.e. chimeric cells or organisms carrying the nucleus of one species and the mitochondrion of another one, typically show an altered respiratory function as compared to native, un-recombined species [4]. More generally, complexes of coadapted genes obviously contribute to developmental stability and fitness [5, 6], and their disruption can result in inbreeding depression, observed in natural hybrid zones or experimental crosses [7].

Such long-term interactions obviously occur between molecules within a cell, and between residues within a molecule. Interacting nucleotides or amino acids determine the folding, and ultimately the function, of biological macromolecules. Proteins and RNAs interact with each other either durably, when forming molecular complexes, or transiently, e.g. as ligand and receptor, as documented in the huge bibliography about structural biology and interaction networks. We should, therefore, expect to detect a strong signal of coevolution within and between genes at the molecular level. Somewhat paradoxically, instances of coevolving proteins or amino acids are not abundant in the molecular evolution literature. Several attempts to detect molecular coevolutionary processes have yielded equivocal results [8], and most phylogenetic methods of sequence analysis make the assumption of independently evolving sites, without biasing the results strongly [9]. Why is coevolution not central in typical analyses of molecular variations between species, despite the importance of long-term interactions within and between genes?

In this chapter, we address the issue of coevolution detection at the molecular level. This is an important question for several reasons. First, characterizing coevolutionary patterns between biological components would help understanding the way they interact and work together. Secondly, coevolution relates to the concept of epistasis, i.e. the fact that the fitness effect of a mutation at a particular focal locus depends on the allelic states taken by some other loci. Epistasis plays a major role in population and quantitative genetics, and in our understanding of the importance and evolution of recombination [10]. Finally, the paradox evoked above (low coevolutionary signal despite strong interactions), if confirmed, would require an explanation. We first present the main ideas of existing methods of molecular coevolution detection, then we review the major biological results obtained in this area over the past 20 years.

Methods of Coevolution Detection

Correlated Patterns

The earliest and simplest methods for detecting coevolution seek correlated patterns across a set of species: two variables are said to be coevolving if they significantly depart the independence assumption, i.e. if the two vectors formed by the values they take in various species are correlated (see [11] for a review). For instance, the two nucleotide sites S_1 and S_2 in figure 1 show correlated patterns since only two pairs of states, A–T and G–C, are observed, whereas the independence hypothesis would predict the occurrence of pairs A–C and G–T as well. This approach is naive in neglecting the underlying phylogenetic correlation: several species can share a common pair of states simply

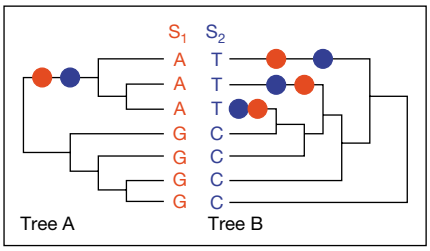


Fig. 1. Accounting for phylogeny when assessing coevolution. Two nucleotide sites, S_1 and S_2, show correlated patterns: from the observed base frequencies at each site (43% A, 57% G for S_1; 43% T, 57% C for S_2) one would expect to observe pairs AT, AC, GT and GC in frequencies 18%, 24%, 24% and 32%, respectively, while they are actually found in frequencies 43%, 0%, 0%, 57%. This rationale neglects the phylogenetic correlation. When phylogeny is taken into account, the coevolution signal can be weak (tree A, a single change at each site) or strong (tree B, three concomitant changes) depending on the tree topology.

because they have recently inherited it from their common ancestor, even in case of independent evolution. Obviously, the coevolutionary signal in figure 1 is stronger if the true phylogeny is tree B, implying three independent co-occurrences of changes, than if it is tree A, where a single change occurred simultaneously at the two sites. The main problem with methods relying on correlated patterns is, therefore, a high type I error, i.e. a high rate of false positive recovery, as in the case of figure 1, tree A. This criticism also applies to early methods of detection of coevolution between genes, based on the correlation between distance matrices (see below): two recently diverged genomes will tend to carry similar proteins, irrespective of their interaction status. Removing the phylogenetic correlation is the main problem of every method of coevolution detection.

Felsenstein first addressed this issue by introducing the ‘independent contrasts’ method, aimed at correcting for phylogeny when correlating quantitative variables [12] and various developments have followed in the so-called ‘comparative analysis’ field (see for instance [13–18]). In the molecular case, where discrete variables are typically considered, several attempts have been made to account for the phylogenetic correlation (see for instance [19–22]).

The vast majority of these methods assume that the underlying phylogeny is known, which is rarely true. How to deal with phylogenetic uncertainties is one of the methodological challenges of comparative analysis. In the case of molecular data, one can simply reconstruct the phylogeny from the analyzed data set. It should be noted that this is a conservative approach as far as coevolution detection is concerned, because tree-building algorithms essentially minimize the number of evolutionary convergences, which can contribute a

substantial fraction of the coevolutionary signal. Assume for instance that tree B (fig. 1) is the true tree. If many coevolving sites follow the pattern of the two sites shown in figure 1, then tree-building algorithms will tend to group the top three species into a single clade, supporting tree A, and decreasing the power to detect coevolution. This is a general consideration about coevolution detection methods, which is relevant to the next section as well.

Correlated Processes

Alternatively to a phylogenetic correction of correlations between observed patterns, one could think of directly comparing the evolutionary histories followed by candidate coevolving entities, therefore incorporating phylogeny in our description of the patterns to be correlated. Unfortunately, these histories are typically unknown, and have to be inferred. Various methods are available to achieve this aim.

Perhaps the most natural approach for detecting coevolving sites in a molecule is to focus on the location of changes in the underlying phylogenetic tree. Obviously, measuring coevolution would be easier if we could observe the scenarios described in figure 1. Several methods based on the phylogenetic mapping of substitutions have been proposed. They essentially differ in (i) how they reconstruct substitution maps, (ii) how they measure correlation between two (or several) substitution maps, and (iii) how they assess its significance.

The substitution mapping (step (i)) can be achieved by maximum-parsimony [8, 23, 24], or preferably in a probabilistic way, using a given substitution model, and integrating over the uncertainty on ancestral states [8, 25, 26]. Step (ii) can be achieved with or without reference to the biochemical nature of substitutions [27]. Dutheil et al. [25] simply measure the correlation coefficient between substitution vectors – entry j in vector $\mathbf{V}_i$ corresponds to the estimated number of changes having occurred in branch j for site i, whereas other studies aim at focusing on compensatory changes, explicitly modeled from our knowledge of amino acid biochemical properties [24, 26]. Step (iii), finally, is typically achieved using simulations [25] or Bayesian posterior predictive probabilities [26]. The heterogeneity of evolutionary rates between sites is an important source of noise: pairs of fast-evolving sites tend to show a higher level of correlation than slowly-evolving ones because the phylogenetic correlation is stronger [25]. The Bayesian approach [26] has the merit of accounting for (integrating over) the uncertainty on tree topology and branch lengths.

Pollock et al. [28], following Pagel [14], proposed an alternative approach to mapping methods in which the departure from the independence assumption is explicit in the model they fit to the data. The method considers the process of evolution of pairs of sites, so that if g is the alphabet size (e.g. four for nucleotides, twenty for amino acids), a substitution matrix of dimension g^2 is

defined, assigning a rate to every change between pairs of states [29]. Under the independence assumption, this $g^2 \times g^2$ matrix is deductible from the standard $g \times g$ one, the equilibrium expected frequency of any (X,Y) pair of states, π_{XY}, being equal to the product of the individual frequencies of states X (π_X) and Y (π_Y). In the Pollock et al. approach, the π_{XY} frequencies are free parameters, which can differ from the $\pi_X \cdot \pi_Y$ products, representing the fact that some pairs of states are more stable, and favored by natural selection, than others. The significance of the coevolutionary signal for a given pair (group) of sites can be assessed through likelihood ratio tests. This is probably the most elegant method proposed so far, but its main drawback is that the number of parameters to be estimated increases dramatically with the number of character states and with the size of candidate site sets. Pollock et al. [28] applied it to pairs of sites only, and recoded the amino acid sequences into a two-state alphabet (large vs. small, positively vs. negatively charged).

Another body of literature addresses the issue of detecting coevolution between proteins, not between sites within a protein, under the assumption that interacting proteins should show correlated evolutionary processes. This is a more complex problem. First, interacting genes can duplicate, leading to complex coevolutionary relationship between orthologous and paralogous members of multigene families [30, 31]. The distinction between orthologs and paralogs was not always explicit in the literature [32], somewhat obscuring the methodological debate. Secondly, these data involve a strong, spurious phylogenetic correlation, because the substitution process of a whole protein is largely influenced by sites evolving independently of the molecular interaction. If two non-interacting proteins evolved in a clock-like manner (i.e., accumulating changes at a constant rate in time), their phylogenetic history would reflect the times between speciation events, and be strongly correlated although they evolve independently (fig. 2, left). The methods of coevolution detection therefore rely on the hypothesis that proteins depart the molecular clock assumption, and aim at seeking pairs of proteins departing it in a correlated way (fig. 2, right).

Most existing methods first summarize the data (aligned sequences) into pairwise distance matrices. The so-called ‘mirror tree’ method [32–34] simply calculates the correlation coefficient between distance matrices (considered as vectors), neglecting the problem of phylogenetic correlation. The obvious problem is to assess the significance of this correlation. An empirical threshold of 0.8 was proposed based on known examples of interacting proteins [33], but this value is probably not universal. More recently, several studies tried to improve the method by taking the species phylogeny into account [35, 36]. A ‘true’ distance matrix is obtained, either from an external source (typically the rRNA tree for bacteria) or by averaging over the set of analyzed proteins. Then the observed distance matrices are corrected by substracting (or projecting onto)

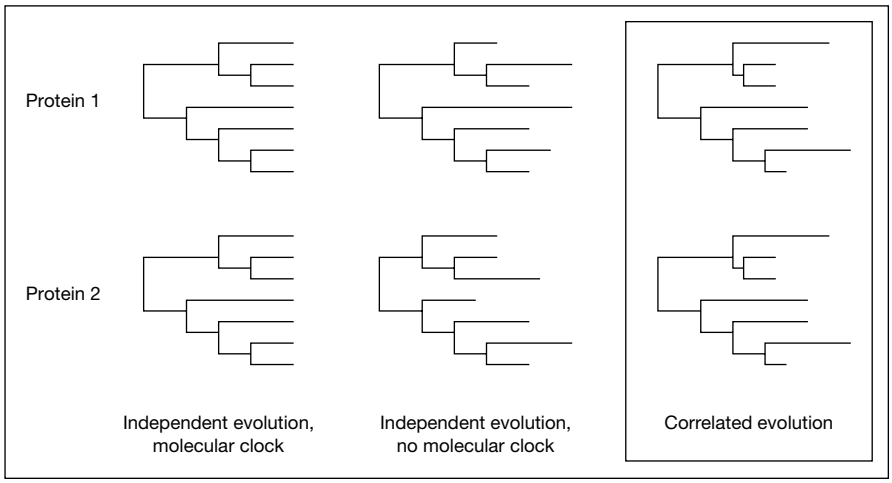


Fig. 2. Correlated phylogenetic patterns. The 'mirror-tree' strategy for detecting co-evolving proteins seeks for correlated tree shapes. This is possible only for proteins departing the molecular clock assumption of evolutionary rate constancy: clockwise evolution results in correlated trees even for non-interacting proteins (left). When proteins do not evolve in a clock-like manner, fast- and slow-evolving lineages will be independent between non-interacting proteins (middle), but correlated between interacting proteins (right).

the 'true' distances, and the residual distances are correlated. Marcotte's 'phylogenetic profile' method [37, 38] relies on BLAST scores to calculate distances: two proteins from a given genome are said to be coevolving if the two vectors of maximal BLAST scores (performed on a series of foreign genomes) are correlated. This approach has the merit of accounting for correlated patterns of gene loss, at the cost of ignoring the orthology/paralogy problem.

Incorporating Functional Information

The methods presented above essentially aim at detecting co-occurring changes in a phylogeny. General considerations about the biochemical nature of nucleotides (e.g. Watson-Crick pairing) or amino acids (charge, volume, polarity) can be taken into account, but the specific fitness effects of the mutations occurring in the analyzed molecules are ignored. Recently, several studies attempted to detect compensatory substitutions by making use of functional data. The idea is to start from mutations known to be deleterious in a model species, and to examine the corresponding sites in related species. Kondrashov et al. [39] reported that roughly 10% of pathogenic amino acid substitutions in 32 human proteins occur in the deleterious state in at least one nonhuman species, and must

therefore have been compensated by other changes – the so-called Dobzhansky-Muller incompatibilities. The combination of phylogenetic and structural information further allowed the identification of candidate compensatory substitutions [39, 40]. A similar result was reported in *Drosophila* [41].

Lessons from Molecular Coevolution Studies

Within Genes: RNAs

The comparative analysis has rapidly proved to be useful for coevolution analysis and was applied with success to structure prediction of ribosomal RNA (rRNA) [11]. RNA has a secondary structure made of one major motif, the double-stranded stem, separated by single stranded regions (loops). Stems have a Watson-Crick like structure, hence relying on A–U and G–C pairs, other pairs leading to mismatches, variably counter-selected depending on the pair. Stem pairs are hence expected to coevolve, since an A to G mutation on one strand may be compensated by a U to C mutation on the opposite strand. On the other hand, if two sites show such a pattern, this can be considered as the signature of a stem pair. The success of this approach, recently confirmed by the structural determination of the ribosome [11] provides evidence of strong, pairwise coevolution between sites within structural pairs. The increasing number of rRNA sequences also allowed to determine new structural motifs involved in higher order structure.

These results raise the question of the underlying mutation mechanism. Coevolving sites show an A–U and G–C pattern series, implying a double-mutation event, which is unlikely. It is now widely accepted that the G–U pair is a less deleterious intermediate, and models have been proposed to put this into account [42]. On a larger evolutionary scale however, such a mechanism failed to explain all observed patterns [25]. More theoretical work is hence required to understand how rRNAs (co-)evolve.

Within Genes: Proteins

The first studies on protein coevolution applied the methodology developed for RNA. This consists of looking for significantly correlated site patterns/processes (see above) in an alignment and using structural information – if available – to check the predictions [21, 22, 27, 43–45]. To increase the power of the method, several authors advocated accounting of the biochemical properties of amino acids, especially volume and charge, for which coevolution is expected. This can be achieved by reducing the proteic alphabet, i.e. grouping amino acids according to their properties [8, 27, 28], or by using a chemical distance to weight comparisons [19, 27]. The most striking result of these studies is the paucity of significantly coevolving groups: among the 15 protein families studied by Tufféry

and Darlu [8], only 6 show a significant coevolution signal, and among the 544 alignments analyzed by Tillier and Lui [21], only 75 have at least one coevolving group. Predicted groups were compared to structural data by checking whether the predicted coevolving sites (i) are close to each other, (ii) belong to known functional regions, like active site or binding sites [22, 27], (iii) are under positive selection [22], or (iv) are known to be crucial for the protein function [27, 45]. Due to the weakness of the coevolution signal, several authors used a different approach and tested explicit coevolutionary hypotheses by starting from available protein structures. They categorized groups of residues (mostly pairs) into groups for which coevolution was expected vs. groups for which it was not (e.g., close vs. distant [19, 28], involved vs. not involved in domain-domain interaction [26]). The coevolutionary signal was then compared between groups.

Since most studies are based on relatively small datasets, it is difficult to draw general conclusions on when and why coevolution occurs within proteins. The percentage of false positives, furthermore, can be high; not every detected pair/group makes sense from a structural viewpoint. However, the following trends were observed:

- Coevolving sites tend to be in close proximity in the structure [21–23, 28, 43, 45]. A possible explanation lies in the local structure arrangement hypothesis, invoked by Gloor et al. [45] and Fares and Travers [22]. The most probable mechanism is volume compensation, although this signal is not always apparent (see for instance [19]).
- Exposed sites are more likely to coevolve than buried sites [28, 44, 45]. This may be because exposed sites are generally less constrained and may be involved in protein-protein interactions, or ligand binding interfaces. Pollock et al. [28] also suggested that sites may coevolve to avoid polymerization of the protein in vivo.
- Some sites within structural motifs do coevolve. Pollock et al. [28] and Dimmic et al. [26] noted that coevolving residues tend to localize in helix ends. This may suggest a role in capping or termination of helices [45]. It was also noted that coevolving residues tend to have a primary distance of 3 or 4, a value which is consistent with the 3.6 residues per turn periodicity of the alpha helix [28]. Some coevolving residues implied in functional sites have been documented, e.g. in the pore domain of the voltage dependent potassium channels [27], the binding site of the methionine amino peptidase 1 [45], in the hinge region of the phosphoglycerate kinase [26], in the HIV gag enzyme [22], and in G protein receptor families [46].

Choi et al. [47] recently took a genome-wide approach by analyzing all proteins with known structure present in the human, rat and dog genomes, extracting meaningful information despite the poor taxonomic sampling. Their results suggest that the main mechanism of coevolution is ionic interaction,

followed by hydrophobic interaction and side chain–side chain hydrogen bond. Surprisingly, they also show that buried sites are more likely to coevolve than exposed sites, contradicting previous works. The authors also measured the coevolution between sites involved in several secondary motifs, and report evidence for significant excess of coevolution between helix and helix, helix and strand, strand and strand, helix and loop, and strand and turn. Such large-scale analyses are probably the next step towards a general understanding of the coevolutionary processes in proteins.

Between Genes

Analyses of protein coevolution were first performed on well-documented interacting proteins for validating the methods [32]. Then the approach was used either on specific example data sets [48], or as a tool for the functional annotation of genomes, with the idea that proteins sharing a common phylogenetic pattern should share some functional characteristics [37, 38]. Such analyses generally confirmed that proteins known to be interacting typically show a higher phylogenetic correlation than random protein pairs, and yielded predictions of yet unknown interacting pairs, which, in many instances, share one or several common Gene Ontology keywords, suggesting a true biological relationship [36]. A separated body of literature makes use of coevolution methods to analyze two families of proteins, each made of many paralogs, globally known to be interacting – e.g. ligands and receptors – with the goal of specifying which pairs of proteins actually interact [30, 31, 49]. It should be noted that these analyses target proteins interacting *sensu lato*, i.e., involved in a common metabolic pathway on which functional constraints (and therefore evolutionary rate) may vary between species. Physical protein-protein interaction is not necessary for such a pattern to appear.

Discussion

A large number of methods have been proposed to detect molecular coevolution, the main challenge being to separate the functional from the phylogenetic correlation. The within-gene problem has been addressed in several distinct ways, from early correlation analyses to sophisticated probabilistic modeling [25, 26, 28], and we would not predict significant methodological breakthrough in this area in the future. One weakness of existing methods, however, is that they typically focus on pairs of sites, while coevolution could occur between groups of three or more sites. Most approaches can easily be extended to deal with an arbitrary number of sites, but the number of distinct subsets of sites is too large to allow an exhaustive examination. Algorithms for selecting

candidate site sets of various sizes are therefore required (Dutheil and Galtier, in preparation). Methods for between-gene coevolution detection, in contrast, appear to be still in their infancy. First, they rely on pairwise distance matrices, which are non-optimal, redundant descriptors of tree shapes. Secondly, the way they correct for the phylogenetic correlation can probably be improved. We foresee that, with the availability of thousands of genes in hundreds of species, the between-gene problem will become closer to the within-gene one (i.e., seeking few interacting pairs out of a large amount of genes), and might benefit from the methodological and bioinformatic developments in this field.

Coevolutionary analyses of molecular data have yielded contrasting patterns: RNA data show a strong coevolutionary signal, especially between Watson-Crick pairs of nucleotides within stems, whereas proteins show only indirect, fuzzy evidence for coevolution between sites. This sounds paradoxical since molecular interactions are obviously of primary importance for protein folding and function. The detection of a substantial amount of Dobzhansky-Muller incompatibilities in human [39] and *Drosophila* [41] further indicate that compensatory changes are common in proteins.

The main difference between RNA and protein data is the tight, long-term pairing existing in RNA stems. Nucleotides in RNA stems bind in a strictly pairwise way, so that a change at one site can only be compensated by a change of its interactor. The pairing, furthermore, is quite conserved throughout evolution, resulting in a strong coevolutionary pattern. Interactions between amino acids within proteins obviously do not obey this scheme. A given residue can interact with several other amino acids, and many distinct kinds of interactions, from loose hydrophobic interactions to tight hydrogen bounds, are possible. A perturbation can therefore probably be compensated in several ways, so that not the same pairs/groups of amino acids will coevolve in the long run. This might explain why protein data sets globally depart the assumption of independent evolution, but rarely allow the detection of well-defined groups of coevolving residues.

This explanation, if true, leaves little hope to approach the interactions between amino acid sites through a coevolutionary analysis: within-molecule epistasis is perhaps strong, but it does not imply correlated phylogenetic patterns if the epistatic relationships evolve quickly. This stresses the need for population genomics studies of coevolution/epistasis, e.g. through the examination of the population fate of double mutants. It might also be the case that the coevolutionary signal in proteins is weak because it has not been sought at the proper spatial scale. Current literature includes studies at the amino acid level, or for whole proteins. Intermediate scales (e.g. coevolution between protein domains) might prove more appropriate. Exploring this dimension will probably require a convergence between the existing within-gene and between-gene methods of coevolution detection.

References

1 van Valen L: A new evolutionary law. Evol Theory 1973;1:1–30.
2 Miller GT, Pitnick S: Sperm-female coevolution in *Drosophila*. Science 2002;298:1230–1233.
3 Rowe L, Arnqvist G: Sexually antagonistic coevolution in a mating system: combining experimental and comparative approaches to address evolutionary processes. Evolution 2002;56:754–767.
4 Blier PU, Dufresne F, Burton RS: Natural selection and the evolution of mtDNA-encoded peptides: evidence for intergenomic co-adaptation. Trends Genet 2001;17:400–406.
5 Clarke GM: The genetic basis of developmental stability. 1. Relationships between stability, heterozygosity and genomic coadaptation. Genetica 1993;89:15–23.
6 Neiman M, Linksvayer TA: The conversion of variance and the evolutionary potential of restricted recombination. Heredity 2006;96:111–121.
7 Pélabon C, Carlson ML, Hansen TF, Yoccoz NG, Armbruster WS: Consequences of inter-population crosses on developmental stability and canalization of floral traits in *Dalechampia scandens* (Euphorbiaceae). J Evol Biol 2004;17:19–32.
8 Tufféry P, Darlu P: Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. Mol Biol Evol 2000;17:1753–1759.
9 Galtier N: Sampling properties of the bootstrap support in molecular phylogeny: influence of non-independence among sites. Syst Biol 2004;53:38–46.
10 Feldman MW, Christiansen FB, Brooks LD: Evolution of recombination in a constant environment. Proc Natl Acad Sci USA 1980;77:4838–4841.
11 Gutell RR, Lee JC, Cannone JJ: The accuracy of ribosomal RNA comparative structure models. Curr Opin Struct Biol 2002;12:301–310.
12 Felsenstein J: Phylogenies and the comparative method. Am Nat 1985;125:1–15.
13 Grafen A: The phylogenetic regression. Philos Trans R Soc Lond B Biol Sci 1989;326:119–157.
14 Pagel M: Detecting correlated evolution on phylogenies – a general-method for the comparative-analysis of discrete characters. Proc R Soc Lond B Biol Sci 1994;255:37–45.
15 Martins EP, Hansen TF: Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 1997;149:646–667.
16 Paradis E, Claude J: Analysis of comparative data using generalized estimating equations. J Theor Biol 2002;218:175–185.
17 Huelsenbeck JP, Rannala B: Detecting correlation between characters in a comparative analysis with uncertain phylogeny. Evolution 2003;57:1237–1247.
18 Housworth EA, Martins EP, Lynch M: The phylogenetic mixed model. Am Nat 2004;163:84–96.
19 Neher E: How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci USA 1994;91:98–102.
20 Akmaev VR, Kelley ST, Stormo GD: A phylogenetic approach to RNA structure prediction. Proc Int Conf Intell Syst Mol Biol 1999;:10–17.
21 Tillier ERM, Lui TWH: Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics 2003;19:750–755.
22 Fares MA, Travers SAA: A novel method to detect intra-molecular coevolution: adding a further dimension to selective constraints analyses. Genetics 2006;173:9–23.
23 Shindyalov IN, Kolchanov NA, Sander C: Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 1994;7:349–358.
24 Fukami-Kobayashi K, Schreiber DR, Benner SA: Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. J Mol Biol 2002;319:729–743.
25 Dutheil J, Pupko T, Jean-Marie A, Galtier N: A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 2005;22:1919–1928.
26 Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R: Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics 2005;21(suppl 1):i126–i135.
27 Fleishman SJ, Yifrach O, Ben-Tal N: An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. J Mol Biol 2004;340:307–318.
28 Pollock DD, Taylor WR, Goldman N: Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol 1999;287:187–198.

29 Tillier ER, Collins RA: High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics 1998;148:1993–2002.
30 Ramani AK, Marcotte EM: Exploiting the co-evolution of interacting proteins to discover interaction specificity. J Mol Biol 2003;327:273–284.
31 Gertz J, Elfond G, Shustrova A, Weisinger M, Pellegrini M, et al: Inferring protein interactions from phylogenetic distance matrices. Bioinformatics 2003;19:2039–2045.
32 Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE: Co-evolution of proteins with their interaction partners. J Mol Biol 2000;299:283–293.
33 Pazos F, Valencia A: Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng 2001;14:609–614.
34 Goh C, Cohen FE: Co-evolutionary analysis reveals insights into protein-protein interactions. J Mol Biol 2002;324:177–192.
35 Sato T, Yamanishi Y, Kanehisa M, Toh H: The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 2005;21:3482–3489.
36 Pazos F, Ranea JAG, Juan D, Sternberg MJE: Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol 2005;352:1002–1015.
37 Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 1999;96:4285–4288.
38 Marcotte EM, Xenarios I, van der Bliek AM, Eisenberg D: Localizing proteins in the cell from their phylogenetic profiles. Proc Natl Acad Sci USA 2000;97:12115–12120.
39 Kondrashov AS, Sunyaev S, Kondrashov FA: Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci USA 2002;99:14878–14883.
40 Kern AD, Kondrashov FA: Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. Nat Genet 2004;36:1207–1212.
41 Kulathinal RJ, Bettencourt BR, Hartl DL: Compensated deleterious mutations in insect genomes. Science 2004;306:1553–1554.
42 Rousset F, Pélandakis M, Solignac M: Evolution of compensatory substitutions through G.U intermediate state in *Drosophila* rRNA. Proc Natl Acad Sci USA 1991;88:10032–10036.
43 Göbel U, Sander C, Schneider R, Valencia A: Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
44 Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW: Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol 2000;17:164–178.
45 Gloor GB, Martin LC, Wahl LM, Dunn SD: Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 2005;44:7156–7165.
46 Oliveira L, Paiva ACM, Vriend G: Correlated mutation analyses on very large sequence families. Chembiochem 2002;3:1010–1017.
47 Choi SS, Li W, Lahn BT: Robust signals of coevolution of interacting residues in mammalian proteomes identified by phylogeny-aided structural analysis. Nat Genet 2005;37:1367–1371.
48 Hao W, Golding GB: Asymmetrical evolution of cytochrome bd subunits. J Mol Evol 2006;62:132–142.
49 Tillier ERM, Biro L, Li G, Tillo D: Codep: maximizing co-evolutionary interdependencies to discover interacting proteins. Proteins 2006;63:822–831.

Nicolas Galtier
Université Montpellier 2 – CC64
Place E, Bataillon
34095 Montpellier (France)
Tel. +33 467 14 48 18, Fax +33 467 14 36 10, E-Mail galtier@univ-montp2.fr

Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 13–29

Evolution of Protein-Protein Interaction Network

T. Makino[a,b,c], *T. Gojobori*[a,d]

[a]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Yata, Mishima, [b]Immunotherapy Division, Shizuoka Cancer Center Research Institute, Shimonagakubo, Nagaizumi-cho, Shizuoka, Japan; [c]Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin, Ireland; [d]Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Aomi, Koto-ku, Tokyo, Japan

Abstract

Protein-protein interactions (PPIs) are one of the most important components of biological networks. It is important to understand the evolutionary process of PPIs in order to elucidate how the evolution of biological networks has contributed to diversification of the existent organisms. We focused on the evolutionary rates of proteins involved with PPIs, because it had been shown that for a given protein-coding gene the number of its PPIs in a biological network was one of the important factors in determining the evolutionary rate of the gene. We studied the evolutionary rates of duplicated gene products that were involved with PPIs, reviewing the current situation of this subject. In addition, we focused on how the evolutionary rates of proteins were influenced by the characteristic features of PPIs. We, then, concluded that the evolutionary rates of the proteins in the PPI networks were strongly influenced by their PPI partners. Finally, we emphasized that evolutionary considerations of the PPI proteins were very important for understanding the building up of the current PPI networks.

Copyright © 2007 S. Karger AG, Basel

Protein-Protein Interaction Network as a Typical Example of Biological Networks

Interactions between proteins and various molecules including proteins themselves are absolutely necessary for sustaining life as a whole. For example, cells are controlled by interacting proteins in metabolic and signaling pathways, such as the molecular machines that replicate, translate and transcribe genes, and build up cell structures. We can classify the biological networks consisting

of such various interactions basically into five types according to the molecules interacting with proteins.

(i) Protein-chemical compound interaction: In the metabolic network, some proteins interact with low-molecular chemical compounds. For example, galactose is metabolized through a series of steps involving the enzymes that are encoded by *GAL1, GAL5, GAL7*, and *GAL10* [1]. These enzymes interact with the appropriate metabolic products.

(ii) Protein-DNA interaction: In the regulatory network, transcriptional factors interact with DNA segments such as the promoter region for transcriptional regulation. For example, genes involved in the galactose metabolism are regulated by the transcriptional factors encoded by *GAL3, GAL4*, and *GAL80*. They interact with the appropriate upstream regions of open reading frames in the DNA.

(iii) Protein-RNA interaction: For the interactions between proteins and nucleotides, proteins interact with not only DNA but also RNA. For example, proteins in the ribosomes in the translation machinery interact with messenger RNAs.

(iv) Protein-lipid interaction: There are proteins interacting with lipids such as phosphoinositides. The phosphoinositides serve as the second messengers that regulate diverse cellular processes [2, 3]. For example, steroid hormone receptors that are transcriptional factors interact with steroid hormones for the transcriptional regulation of target genes [4].

(v) Protein-protein interaction: Finally, protein-protein interactions (PPIs) are well-studied components of biological networks. PPIs are involved in a number of biological processes such as protein transportation and degradation, cell cycle progression, polarity, gene expression and DNA repair. For example, the transcriptional factor encoded by *GAL80* as already mentioned above interacts with the other transcriptional factors encoded by *GAL3* and *GAL4* for the regulation of galactose utilization.

Recently, global studies on PPIs have been investigated not only in prokaryotes, which are *Helicobacter pylori* [5] and *Escherichia coli* [6], but also in eukaryotes, which are *Plasmodium falciparum* [7], *Caenorhabditis elegans* [8], *Drosophila melanogaster* [9, 10] and human [11, 12]. In particular, *Saccharomyces cerevisiae* provides a great advantage for the study of PPIs, because a vast amount of information about PPIs has been produced not only by hundreds of small-scale experiments but also by the high-throughput yeast two-hybrid system (Y2H system; [13, 14]) and mass spectrometry of coimmunoprecipitated protein complexes (Co-IP; [15–17]). However, the high-throughput data on PPIs are known to contain a number of false-negative and false-positive interactions. In the case of the false-negative interactions, the PPIs sometimes could not be detected in the Y2H system for full-length ORFs. This is because

full-length proteins often show much weaker signals than appropriately trimmed protein regions containing interacting proteins. On the other hand, proteins having low expression levels will not be able to be identified by the Co-IP, because of the limitation of the sensitivity for the system. Therefore, the detection of the PPIs should be conducted by the both methods that are mutually complementary. Proteins such as transcriptional factors activate the expression of a reporter gene in the Y2H system and lead to false-positive interactions. Contaminant proteins with high expression levels tend to be recovered in coimmunoprecipitated protein complexes (Co-IP), even if they do not actually interact with one another. Consequently, the high-throughput data require further examination for their accuracy. Several methods for removing dubious PPIs from the original data were developed, and as a result, the credible PPIs have become enriched [18–20].

Evolutionary Studies of Protein-Protein Interaction Networks

Until now, molecular evolutionary analyses have mainly focused on individual genes regardless of how they are involved with the interactions among their gene products. However, it is interesting to carry out evolutionary analyses of a group of genes in which the encoding proteins interact with one another in the PPI network. In these analyses, it is important to examine how selective pressures affect gene products as the components of PPI networks. It is of particular interest to study how the organization of proteins as members of PPI network affects the evolutionary rates of their corresponding genes.

It should be noted that duplicated genes encoding proteins in PPI networks provide us with a unique opportunity of making fair comparisons of the genes under the same initial condition. The pair of proteins encoded by a duplicated gene pair often share PPI partners [20, 21], although some of the PPI partners may be lost later in the evolutionary process. In fact, there are a lot of duplicated pairs encoding proteins not having the shared PPI partners [21]. Therefore, we examined the relationship in the evolutionary rates between a duplicated protein in PPIs and its counterpart ('*Differential evolutionary rates of duplicated genes in protein interaction network*' in this chapter; [22]).

It has been shown that proteins sharing functions tend to interact in the PPI networks [23, 24]. There is a strong correlation between the structure of the PPI network and the functions of the proteins in the network [25]. In other words, many functions appear to be particular parts in the PPI networks. On the other hand, the recent study gave us an interesting insight [26]. The authors have shown that there are many proteins interacting with their PPI partners having different functions. For example, mitogen-activated protein kinase (MAPK)

interacts with proteins having different functions that are involved in ribosomal biogenesis, cytoskeleton and directional cell growth. In particular, it has been shown that such PPIs have biological importance according to an experiment of double gene deletion for genes encoding the protein and its PPI partner. PPIs are not in a uniform state as mentioned above. It is of great interest to study how the interacting proteins have been evolutionarily influenced by their PPI partners in the PPI network. Therefore, we examined the differences in evolutionary rate among the interacting proteins involved in different PPIs ('*The evolutionary rate of a protein is influenced by features of the interacting partners*' in this chapter; [27]).

Differential Evolutionary Rates of Duplicated Genes in Protein Interaction Network

The functional constraints of proteins involved in the PPI network are composed of several factors. The so-called fitness effects as well as the gene expression level are typical factors, because they are known to be negatively correlated with the rate of amino acid substitutions [28–31]. The number of PPIs for a given protein is also an important factor for determining its evolutionary rate. It has been reported that the number of PPI partners for proteins is negatively correlated with their evolutionary rates [32, 33]. Therefore, after gene duplication, the differentiation of PPIs through the PPI losses and/or PPI gains during evolution may affect the evolutionary rates of duplicated pairs. For a duplicated gene pair, it has been shown that one copy usually has more PPI partners than the other [34].

Gene duplication is one of the major evolutionary mechanisms for generating novel genes [35]. After gene duplication, one of the pair may be redundant, such that functional constraint is relaxed to allow one or both to differentiate as long as the original function is retained as a whole. Three pathways have been proposed for functional differentiation of duplicated genes [36]. First, one copy may be silenced by accumulation of deleterious mutations and eventually become indistinguishable from the nearby noncoding genomic regions in the absence of functional constraints, while the other copy retains the original function. Second, while one copy maintains the original function, the other acquires a novel function possibly by advantageous mutations. Third, both copies accumulate mutations that alter the original function, but compensate for the original function cooperatively. When a duplicated gene pair functionally differentiates, the evolutionary rate may be accelerated in one or both due to the relaxation of negative selection or the enhancement of positive selection [37]. In yeast, it has been proposed that the differentiation process is asymmetrical rather than symmetrical to minimize the risk of deleterious mutations [34]. It is

therefore expected that the acceleration of evolutionary rates occurs mainly in one of two copies after gene duplication. However, it is not yet known how the duplicated gene products affect their PPIs in evolution.

Duplicated products often interact with the same proteins [20]. One proposed model for the losses and/or gains of PPIs provides the reason why the products of a duplicated gene pair often share PPI partners [21]. In this model, although some duplicated pairs lose PPIs during the evolutionary process, many duplicated pairs retain some shared PPI partners. In a recent study, the magnitude of functional divergence for duplicated pairs was measured by using the number of shared PPI partners between all pairs in the PPI networks [38]. To examine the relationship between the evolutionary rate and the functional differentiation of duplicated gene products, we focused on the shared PPI partners that were considered to represent characteristics of the functional differentiation of the duplicated gene products, because the products sharing PPI partners would not have largely diverged.

The purpose of the study is to understand how gene duplication influences the evolution of PPI networks. To study the relationship between gene duplication and the evolutionary rates of the gene products with PPI partners, we used the PPIs in *Saccharomyces cerevisiae* that have well been documented based not only on hundreds of small-scale experiments but also on high throughout methods. We set up and examined the hypothesis that the ratios of evolutionary rates (faster rate/slower rate) for the pairs sharing any PPI partners are lower than those for the pairs sharing no PPI partners. We then discuss the mechanisms of the functional differentiation after gene duplication on the basis of the results obtained.

Losses of PPIs for Proteins Encoded by Duplicated Genes

Soon after gene duplication, the protein encoded by one copy should interact with the same set of proteins as the other, because both proteins are identical. It has been proposed that PPI partners of proteins encoded by duplicated genes change through PPI losses or PPI gains during evolution [21]. For a duplicated gene pair, it has been shown that one copy usually has more PPI partners than the other [34]. However, it was unclear which of the two mechanisms, namely PPI losses and PPI gains, is the major force of the evolution of PPIs. Proteins under strong functional constraints would be hard to change their PPI partners during evolution, because they are conservative. The PPI losses of the proteins may accelerate their evolutionary rates, because it has been reported that the evolutionary rate is negatively correlated with the number of PPIs [32, 33]. If the PPI losses occur more often than the PPI gains for a duplicated pair, the protein encoded by one copy evolving at a slower rate would have more PPI partners than the other.

To examine this possibility, we used duplicated pairs generated by genome duplication in *Saccharomyces cerevisiae*, which occurred about 100 million years ago [39, 40]. For each pair of gene products, we examined whether the protein with more PPI partners evolved more slowly than the other with less partners. We then found that a protein with more PPI partners evolved at a slower rate in 134 (62%) out of the 216 pairs examined, which was significantly greater than expected under the null hypothesis of random association between the number of PPI partners and the evolutionary rate (50%). We found that the protein encoded by one copy evolving at a slower rate had more PPI partners than the other copy. The results indicated that the PPI losses have occurred more often than the PPI gains for a copy evolving at a faster rate, on the assumption that PPIs of a copy evolving at a slower rate are conservative in the evolutionary process.

Functional Divergence through Changes in PPIs

After gene duplication, there are at least two possible pathways for PPI divergence of the proteins encoded by a duplicated gene pair. First, one encoded by a duplicated pair keeps the shared PPI partners, and the other loses all the shared PPI partners. The evolutionary rate of the former would be slower than that of the latter, because the former has to maintain the original function while the latter is free from it. In other words, they are likely to evolve at different rates. Second, both proteins share some of the PPI partners. In this case, both proteins will still have similar functions, and their sequences would not change by mutations as drastically as in the latter of the first case. The evolutionary rates of the gene products sharing PPI partners may not significantly differ from one another. If duplicated gene products lose the shared PPI partners, the ratio of evolutionary rates for the pair (faster rate/slower rate) may be higher than that for functionally similar pairs.

To test this hypothesis, we examined whether F_1/S_1 were higher than F_2/S_2, where F and S denote faster rate and slower rate, respectively, and subscripts 1 and 2 refer to the cases of sharing no PPIs and sharing PPIs, respectively (fig. 1). Here we defined duplicated pairs sharing PPIs as the pairs sharing at least one PPI partner. There were 124 duplicated pairs sharing no PPI partners and 130 duplicated pairs interacting with one another or sharing PPI partners. F_1/S_1 was significantly higher than F_2/S_2 (fig. 2).

For a duplicated gene pair, if the protein encoded by one copy evolving at a faster rate has not been silenced during evolution, it would have lost its PPI partners and have a chance of finding a new PPI partner under the weak or no functional constraints. On the other hand, the PPIs for the protein encoded by one copy evolving at a slower rate would be conservative with relatively strong functional constraints. For duplicated pairs, the gene product evolving at a

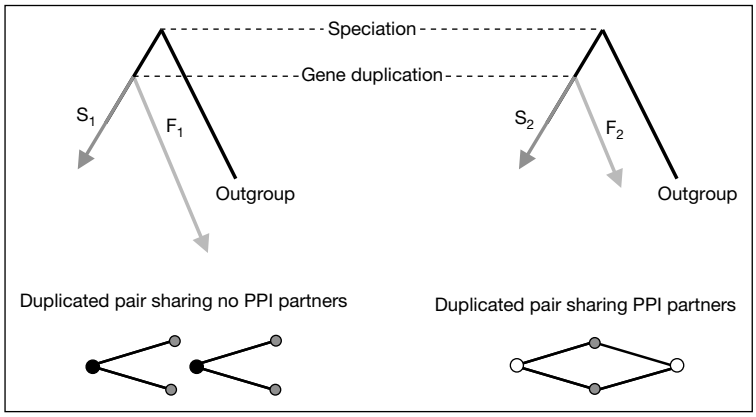


Fig. 1. Schematic representations of F_1, S_1, F_2, and S_2. Closed circles and open circles respectively mean proteins encoded by duplicated gene pair sharing no PPI and sharing PPIs. F (light gray arrow) and S (gray arrow) mean faster rate and slower rate, respectively, and subscripts 1 and 2 refer to the cases of sharing no PPI and sharing PPIs for duplicated pairs, respectively. The ratio of evolutionary rates for duplicated pairs after gene duplication was estimated by a faster evolutionary rate of one copy/a slower rate of another copy (F_1/S_1; F_2/S_2).

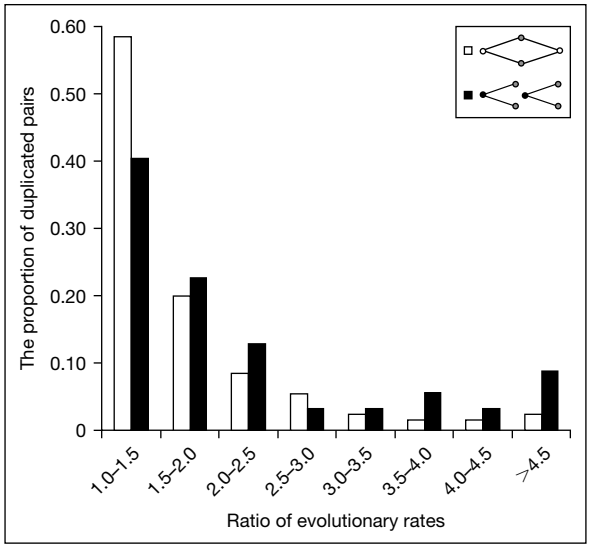


Fig. 2. Ratios of evolutionary rates for duplicated pairs sharing PPI partners and sharing no PPI partners. Open bars indicate duplicated pairs interacting with one another or sharing PPI partners, while closed bars indicate duplicated pairs sharing no PPI partners.

Table 1. Results of relative rate test for duplicated pairs having PPI partners and sharing no PPI partners in functional class 'transcription'

	Number of duplicated pairs	
	sharing PPI partners	not sharing PPI partners
Significant difference of rates	10	19
No significant difference of rates	13	7

faster rate will lose the shared PPI partners more frequently than the other. This implies that a pair of proteins encoded by a duplicated gene pair having few shared PPI partners evolves at different rates. In fact, the present study indicates that pairs sharing no PPI partners show a larger ratio of evolutionary rates than those sharing PPI partners, although it has been reported that a simple relationship between sequence divergence and their functional divergence revealed by the PPI network analysis could not be established [38]. When a duplicated gene pair shares no PPI partners, it is possible that the gene products interact with different PPI partners with different functions. This means that gene duplication will lead to the functional differentiation of the duplicated gene products through the PPI losses and/or PPI gains, which will then cause a change in their evolutionary rates.

Tendency of PPI Divergence for Duplicated Pair in Different Functional Classes

For investigating the functions of duplicated gene products, we used functional classification established by the MIPS database [41]. In the functional class of 'transcription', there were significantly many duplicated pairs sharing no PPI partners and having significant difference in evolutionary rates (table 1). There were also statistically significant differences in the rate between the two copies in the functional class of 'protein fate' (table 2). These results indicate that the PPIs of the proteins included in these functional classes tend not to be conservative in the evolutionary process, resulting in a change in their evolutionary rates. The other functional classes showed no significant difference in ratio of evolutionary rates between duplicated pairs sharing PPI partners and those sharing no PPI partners.

We found many cases of pairs sharing no PPI partners in the functional classes such as 'transcription' and 'protein fate'. For example, YNR023W and YCR052W (a duplicated pair in 'transcription') do not share PPI partners, and

Table 2. Results of relative rate test for duplicated pairs having PPI partners and sharing no PPI partners in functional class 'protein fate'

	Number of duplicated pairs	
	sharing PPI partners	not sharing PPI partners
Significant difference of rates	2	11
No significant difference of rates	10	5

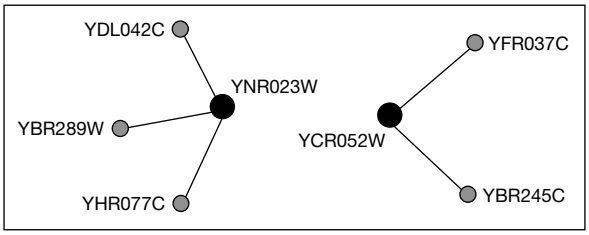


Fig. 3. An example for the pair of proteins encoded by duplicated gene pairs and their PPI partners. The circles and lines represent proteins and PPIs, respectively. The circles in gray are PPI partners. The closed circles represent proteins encoded by the duplicated gene pair (YNR023W and YCR052W), which are a subunit of SWI/SNF global transcription activator complex and a subunit of the RSC chromatin-remodeling complex, respectively.

have a significant difference in evolutionary rate between them. In addition, they are subunits in different protein complexes. YNR023W is a subunit of SWI/SNF global transcription activator complex, and YCR052W is a subunit of the RSC chromatin-remodeling complex (fig. 3; [42]). We consider the significant difference in evolutionary rate between the two copies is caused by drastic changes in the PPI partners during evolution. Although the proteins encoded by these duplicated gene pairs would have interacted with the same PPI partners immediately after the gene duplication, one of the copies would have subsequently changed its PPI partners and diverged its functions. It is thus suggested that YCR052W, which evolves at a faster rate than YNR023W, would have obtained novel functions by changing their PPI partners. Thus, the evolutionary comparison of the PPI partners of one copy in a duplicated pair with those of the other is important for understanding their functional differentiations through PPI network divergence.

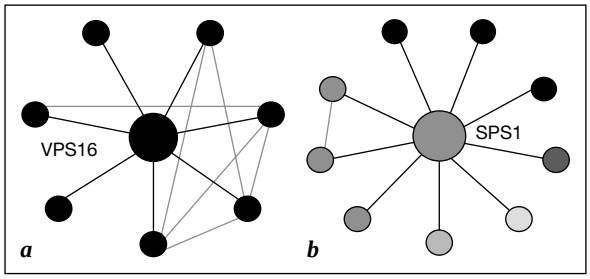


Fig. 4. ***a*** A protein in a functional module and (***b***) a protein in a framework module of the PPI network. The filled circles and lines represent proteins and PPIs, respectively. The black lines indicate interactions between VPS16 and its PPI partners and between SPS1 and its PPI partners. The gray lines indicate interactions among PPI partners. VPS16 interacts with proteins classified into the same functional class 'protein fate' on Munich Information Center for Protein Sequences database [41]. The different grey scales of the circles in ***b*** mean different functional classes. SPS1 interacts with proteins classified into different functional classes 'protein fate', 'cell cycle/DNA processing', 'metabolism', 'cellular transport', and 'transcription', respectively.

The Evolutionary Rate of a Protein is Influenced by Features of the Interacting Partners

When a two-dimensional presentation of PPI networks is made using a node and a line between neighboring nodes as a protein and an interaction between neighboring proteins, respectively, the PPI network is represented by a very complex structure of spider web-like networks. It has been reported, in this type of representation, that there are proteins tightly clustered in a particular part of the PPI network [43]. In particular, the proteins sharing a particular functional class tend to appear in the same part of a PPI network, making a cluster of the so-called 'functional module' in the PPI network [25]. Here, a functional class represents a category into which a group of particular proteins is classified according to the functional definitions. In other words, a functional module of the network is generally defined as a cluster of proteins sharing the same functional class that occupies a specific part of the network. In the PPI networks, the proteins building up a functional module have more interactions to other proteins within the functional module than to those outside the module. For example, VPS16 of *Saccharomyces cerevisiae* is clustered in a functional module that is required for sorting proteins in vacuolare (fig. 4a).

On the other hand, there are proteins known to interact with those having different functional classes [26]. Calmodulin, which is a master regulator of

calcium-mediated signaling [44], interacts with several proteins of different functional classes such as homeostasis of cations, protein folding and stabilization, budding, cell polarity and filament formation [26]. For these proteins, the gene expression patterns do not correlate with those of their PPI partner proteins, suggesting that they interact with the PPI partners at different subcellular localizations or different time points. Let us call these the proteins in a framework module. In other words, the protein in a framework module is defined as a protein mediating different functions by interactions of proteins having different functional classes. For example, SPS1 encoding ser/thr protein kinase of *S. cerevisiae* is in a framework module, and interacts with proteins classified into different functional classes (fig. 4b). Therefore, the number of interactions among the PPI partners of these proteins in the framework module is expected to be smaller than that of the proteins in the functional module.

It is interesting to investigate the extent to which the evolutionary rate of proteins is influenced by the nature of PPIs. Therefore, we examined the differences in evolutionary rate among the proteins having different types of PPI partners. The difference in the evolutionary rate can be interpreted by the difference in functional constraints if the mutation rate does not vary much with the proteins. Thus, we would also discuss the differences in functional constraint among the proteins having different types of PPI partners in the PPI network.

SF vs. DF Proteins

Proteins in the PPI networks would have evolved under the influence of their PPI partners. It has been reported that the number of PPI partners is correlated significantly to their evolutionary rates [32, 33]. A recent study reported that proteins in the center of the PPI networks evolve more slowly, regardless of the number of PPI partners [45]. When the proteins lose or gain their PPI partners during evolution, an allowable degree of their amino acid substitutions may depend not only on the number of their PPI partners but also on the features of their PPI partners. It has been known that proteins sharing the same functional class tend to interact with each other [23, 24]. On the contrary, there are proteins that interact with those belonging to different functional classes [26]. Here, we defined a protein having PPI partners of the same functional class with a high frequency as an SF (the Same Function) protein, on the other hand, a protein having PPI partners of different functional classes with a high frequency as a DF (the Different Function) protein. It is of particular interest to know which of the SF or DF proteins is under stronger functional constraints in the evolutionary process. Therefore, we examined whether the evolutionary rates of the proteins in the PPI network have been strongly influenced by the PPI partners having the same or different functional classes. To answer the question, we compared the evolutionary rates of the SF proteins with those of

the DF proteins in yeast PPI networks. For this comparative study, we used the evolutionary distances for 1,035 SF and 763 DF proteins for the comparison. As a result, we found that the DF proteins evolved at a slower rate, with statistical significance, than the SF proteins. Thus, we concluded that the DF proteins are under much stronger functional constraints than the SF proteins.

DP vs. SP Proteins

It has been reported that there are proteins tightly clustered in a particular part of the PPI network [43]. Denoting proteins in dense and sparse parts of the PPI network as the DP (Dense Part) and SP (Sparse Part) proteins, respectively, we defined them using the clustering coefficient [46]. We examined the differences in evolutionary rates between DP proteins in a dense part of PPI networks and SP proteins in a sparse part of PPI networks. When we compared the evolutionary rates of the 668 DP proteins with those of the 965 SP proteins, we found that the SP proteins evolved at a slower rate, with statistical significance, than the DP proteins. Interestingly enough, this is also opposite to our expectation. Before conducting the present study, we speculated that the DP proteins would have slower rates, because it has been reported that proteins having cohesive patterns of PPIs are more evolutionarily conservative than other proteins in the PPI network [47]. In contrast, our observation suggests that the proteins in a sparse part of the PPI network could be more important than those in a dense part. It is possible that the PPI partners in a sparse part in the PPI network are indispensable because of possible scarceness of substitutable PPI partners. This is an interesting and meaningful finding.

Comparison of Evolutionary Rates among SF-DP, SF-SP, DF-DP and DF-SP Proteins

According to the results described above, we reasonably hypothesized that the DF-SP proteins would evolve at the slowest rate in the proteins examined. To test the hypothesis, we statistically compared the evolutionary rates among the 443 SF-DP, 353 SF-SP, 122 DF-DP and 457 DF-SP proteins. We found that out of all proteins examined the DF-SP proteins evolved certainly at the slowest relative evolutionary rate (fig. 5). The result suggests that the proteins having the PPI partners belonging to different functional classes and being in a sparse part of the PPI network are under the strongest functional constraints, implying that those proteins are possibly important for the maintenance and survival of the PPI network.

We have found that the DF proteins evolved at a slower rate than the SF proteins. The observation suggests that the proteins involved with multi-different biological processes in the PPI network are under strong functional constraints. We have also shown that the SP proteins evolved at a slower rate than the DP

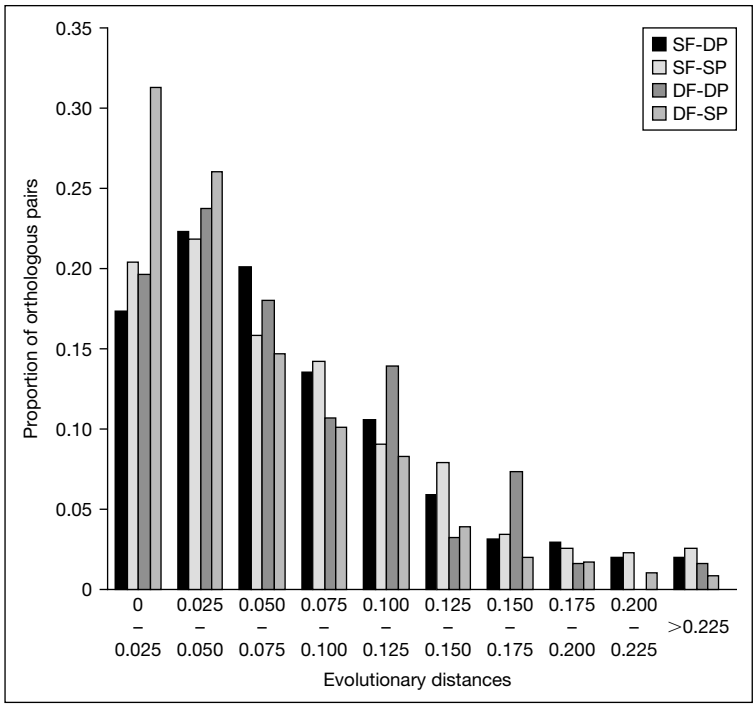


Fig. 5. Distribution of evolutionary distances for the SF-DP, SF-SP, DF-DP, and DF-SP proteins. The evolutionary distance is measured as the number of amino acid substitutions per site.

proteins. In fact, we have shown that the DF-SP proteins evolved at the slowest rate among all interacting proteins studied. This might be explained if loss of function in DF-SP proteins affected multiple biological processes more so than that of proteins with other interaction properties. These results strongly suggest that the evolutionary rates of proteins depend on the nature of interacting proteins in the PPI network.

For the evolutionary studies of proteins in the PPI networks, it has been shown that proteins involved in protein complexes are more evolutionarily conservative than other proteins in the PPI networks [48]. A protein complex can be considered as a typical example of SF proteins, because all the subunits are regarded as belonging to the same functional class due to a particular functional manifestation of the whole protein complex. To confirm this consideration, we compared a proportion of subunits in protein complexes for the SF proteins with that for DF proteins using the protein complex data set in the MIPS database [41]. As expected, we found that the SF proteins contained more subunits

of protein complexes than the DF proteins (data not shown). Although the SF proteins contained relatively many subunits of a protein complex, our results clearly showed that the SF proteins are evolutionarily much less conservative than the DF proteins. Moreover, it has been reported that proteins having cohesive patterns of PPIs are more evolutionarily conservative than other proteins in the PPI network, and tend to be subunits of protein complexes [47]. The proteins would be under strong structural constraints, because many of the proteins are in an extremely dense part of the PPI network. Although the authors particularly showed high evolutionary conservation of the proteins having cohesive patterns of PPIs, our finding is that the DF-SP proteins are under the strongest functional constraints among all interacting proteins studied. This conclusion highlights the importance of studying the evolution of the DF-SP proteins for understanding essential features of PPI network evolution.

Prospect of Studies in PPI Network Evolution

We focused on two themes to study the evolution of protein-protein interaction networks as a typical example of biological networks.

First, we focused on a relationship between the PPI divergences of duplicated gene products and their evolutionary rates, and examined whether the difference in evolutionary rate exists between a duplicated pair of genes encoding proteins involved in PPIs. Our results showed the evolutionary rate of a protein having more PPI partners is much slower than that of the other having fewer PPI partners. Moreover, we found that the ratios for duplicated pairs sharing PPI partners are significantly lower than the ratios for pairs sharing no PPI partners. When a duplicated pair shares no PPI partners, it is possible that the gene products interact with the PPI partners having different functions. These results clearly indicate that gene duplication leads to the functional differentiation of the duplicated gene pairs through PPI losses and/or PPI gains. The functional differentiation would cause eventually the change in their evolutionary rates. The evolutionary comparison of the PPI partners of one copy in a duplicated pair with those of the other copy gives an important clue for understanding their functional differentiations through PPI network divergence.

Second, we focused on the differences in evolutionary rates among interacting proteins having different types of PPI partners, because it is of particular interest to know how the PPIs influence the evolutionary rate, namely the rate of amino acid substitutions. In fact, we showed that the DF proteins, which interact with PPI partners in different functional classes with a high frequency, evolve at a slower rate than the SF proteins do, which interact with PPI partners in the same functional class with a high frequency. It suggests that the

interacting proteins involved in multi-different biological processes would be under strong functional constraints. We also showed that SP proteins, which are in sparse parts of the PPI networks, evolve at a slower rate than the DP proteins, which are in dense parts of the networks. The result indicates that the weaker relationship among PPI partners of proteins is, the more slowly the interacting proteins evolve. These results strongly suggested that the evolutionary features of the interacting proteins have been influenced by the type of their PPIs such as functional and framework modules.

We clearly pointed out the advantage of utilizing a vast amount of information about PPIs in the molecular evolutionary studies of biological networks. In particular, we successfully showed that the evolution of proteins as the components of PPI networks can be understood, to a reasonably great extent, through the evolutionary rates. Finally, we would like to emphasize that this line of studies will give us an important insight into the understanding of evolutionary processes of the PPI networks.

Acknowledgements

This project is, in part, supported by the Genome Network Project of MEXT (Ministry of Education, Culture, Sports, Science and Technology) and BIRC (Biological Information Research Center) at AIST (National Institute of Advanced Industrial Science and Technology).

References

1 Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, et al: Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 2001;292:929–934.
2 Odorizzi G, Babst M, Emr SD: Phosphoinositide signaling and the regulation of membrane trafficking in yeast. Trends Biochem Sci 2000;25:229–235.
3 Wera S, Bergsma JC, Thevelein JM: Phosphoinositides in yeast: genetically tractable signalling. FEMS Yeast Res 2001;1:9–13.
4 Tsai MJ, O'Malley BW: Molecular mechanisms of action of steroid/thyroid receptor superfamily members. Annu Rev Biochem 1994;63:451–486.
5 Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, et al: The protein-protein interaction map of *Helicobacter pylori*. Nature 2001;409:211–215.
6 Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, et al: Interaction network containing conserved and essential protein complexes in *Escherichia coli*. Nature 2005;433:531–537.
7 LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, et al: A protein interaction network of the malaria parasite *Plasmodium falciparum*. Nature 2005;438:103–107.
8 Li S, Armstrong CM, Bertin N, Ge H, Milstein S, et al: A map of the interactome network of the metazoan *C. elegans*. Science 2004;303:540–543.
9 Formstecher E, Aresta S, Collura V, Hamburger A, Meil A, et al: Protein interaction mapping: a *Drosophila* case study. Genome Res 2005;15:376–384.
10 Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al: A protein interaction map of *Drosophila melanogaster*. Science 2003;302:1727–1736.

11 Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al: Towards a proteome-scale map of the human protein-protein interaction network. Nature 2005;437:1173–1178.
12 Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al: A human protein-protein interaction network: a resource for annotating the proteome. Cell 2005;122:957–968.
13 Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, et al: Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci USA 2000;97:1143–1147.
14 Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al: A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 2000;403:623–627.
15 Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002;415:141–147.
16 Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al: Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 2002;415:180–183.
17 Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al: Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 2006;440:637–643
18 Bader GD, Hogue CW: Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol 2002;20:991–997.
19 Bader JS, Chaudhuri A, Rothberg JM, Chant J: Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 2004;22:78–85.
20 Deane CM, Salwinski L, Xenarios I, Eisenberg D: Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics 2002;1:349–356.
21 Wagner A: The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 2001;18:1283–1292.
22 Makino T, Suzuki Y, Gojobori T: Differential evolutionary rates of duplicated genes in protein interaction network. Gene 2006;385:57–63.
23 Ge H, Liu Z, Church GM, Vidal M: Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. Nat Genet 2001;29:482–486.
24 Schwikowski B, Uetz P, Fields S: A network of protein-protein interactions in yeast. Nat Biotechnol 2000;18:1257–1261.
25 Yook SH, Oltvai ZN, Barabasi AL: Functional and topological characterization of protein interaction networks. Proteomics 2004;4:928–942.
26 Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, et al: Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 2004;430:88–93.
27 Makino T, Gojobori T: The evolutionary rate of a protein is influenced by features of the interacting partners. Mol Biol Evol 2006;23:784–789.
28 Hirsh AE, Fraser HB: Protein dispensability and rate of evolution. Nature 2001;411:1046–1049.
29 Jordan IK, Rogozin IB, Wolf YI, Koonin EV: Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res 2002;12:962–968.
30 Pal C, Papp B, Hurst LD: Highly expressed genes in yeast evolve slowly. Genetics 2001;158:927–931.
31 Wilson AC, Carlson SS, White TJ: Biochemical evolution. Annu Rev Biochem 1977;46:573–639.
32 Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW: Evolutionary rate in the protein interaction network. Science 2002;296:750–752.
33 Fraser HB, Wall DP, Hirsh AE: A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol 2003;3:11.
34 Wagner A: Asymmetric functional divergence of duplicate genes in yeast. Mol Biol Evol 2002;19:1760–1768.
35 Ohno S: Evolution by Gene Duplication. Springer, Berlin, 1970.
36 Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: Preservation of duplicate genes by complementary, degenerative mutations. Genetics 1999;151:1531–1545.
37 Li WH, Gojobori T: Rapid evolution of goat and sheep globin genes following gene duplication. Mol Biol Evol 1983;1:94–108.
38 Baudot A, Jacq B, Brun C: A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein-protein interaction network. Genome Biol 2004;5:R76.

39 Kellis M, Birren BW, Lander ES: Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 2004;428:617–624.
40 Wolfe KH, Shields DC: Molecular evidence for an ancient duplication of the entire yeast genome. Nature 1997;387:708–713.
41 Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, et al: MIPS: a database for genomes and protein sequences. Nucleic Acids Res 2002;30:31–34.
42 Cairns BR, Lorch Y, Li Y, Zhang M, Lacomis L, et al: RSC, an essential, abundant chromatin-remodeling complex. Cell 1996;87:1249–1260.
43 Spirin V, Mirny LA: Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA 2003;100:12123–12128.
44 Davis TN, Urdea MS, Masiarz FR, Thorner J: Isolation of the yeast calmodulin gene: calmodulin is an essential protein. Cell 1986;47:423–431.
45 Hahn MW, Kern AD: Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 2005;22:803–806.
46 Watts DJ, Strogatz SH: Collective dynamics of 'small-world' networks. Nature 1998;393:440–442.
47 Wuchty S, Oltvai ZN, Barabasi AL: Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat Genet 2003;35:176–179.
48 Teichmann SA: The constraints protein-protein interactions place on sequence divergence. J Mol Biol 2002;324:399–407.

Takashi Gojobori
Center for Information Biology and DNA Data Bank of Japan
National Institute of Genetics
1111 Yata, Mishima-shi, Shizuoka-ken 411-8540, Japan
Tel. +81-55-981-6847, Fax +81-55-981-6848, E-mail tgojobor@genes.nig.ac.jp

Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 30–47

Bacterial Flagella and Type III Secretion: Case Studies in the Evolution of Complexity

M.J. Pallen[a], *U. Gophna*[b]

[a]University of Birmingham Medical School, Birmingham, United Kingdom;
[b]Department of Molecular Microbiology and Biotechnology, The George S. Wise Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv, Israel

Abstract

Bacterial flagella at first sight appear uniquely sophisticated in structure, so much so that they have even been considered 'irreducibly complex' by the intelligent design movement. However, a more detailed analysis reveals that these remarkable pieces of molecular machinery are the product of processes that are fully compatible with Darwinian evolution. In this chapter we present evidence for such processes, based on a review of experimental studies, molecular phylogeny and microbial genomics. Several processes have played important roles in flagellar evolution: self-assembly of simple repeating subunits, gene duplication with subsequent divergence, recruitment of elements from other systems ('molecular bricolage'), and recombination. We also discuss additional tentative new assignments of homology (FliG with MgtE, FliO with YscJ). In conclusion, rather than providing evidence of intelligent design, flagellar and non-flagellar Type III secretion systems instead provide excellent case studies in the evolution of complex systems from simpler components.

Copyright © 2007 S. Karger AG, Basel

Type III Secretion

Type-III secretion is one of several different forms of protein secretion employed by bacteria to transport proteins from the cytoplasm to the external milieu [1–4]. The systems that mediate this kind of secretion, the type III secretion systems (T3SSs), are exquisitely engineered molecular pumps, harnessing hydrolysis of ATP to drive export of proteins across the bacterial cell envelope. Each T3SS consists of over a dozen different kinds of protein and provides a paradigm of how hierarchical gene regulation, complex protein-protein interactions and controlled protein secretion can result in the assembly of a complex multi-protein structure tightly orchestrated in time and space.

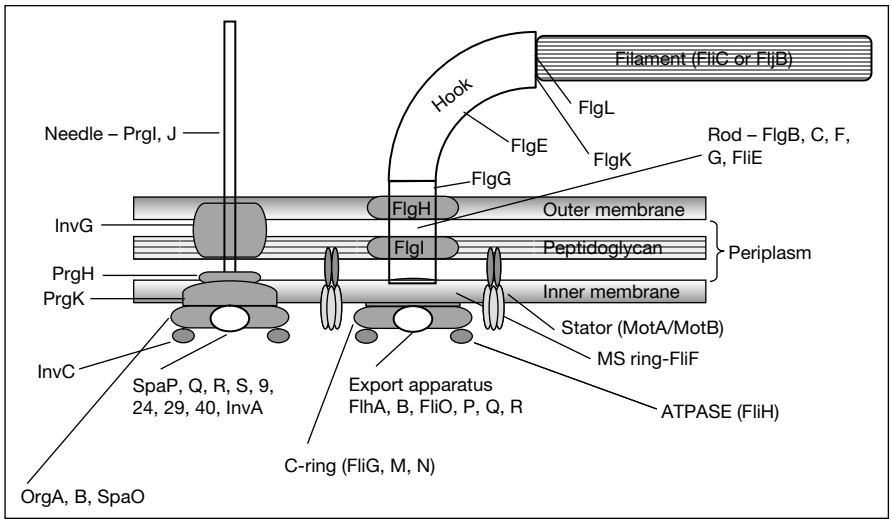


Fig. 1. Flagellar and non-flagellar type III secretion systems.

Type III secretion systems are deployed in two functionally distinct contexts (fig. 1):

- biosynthesis of the bacterial flagellum, the chief organelle of motility in bacteria (mediated by the flagellar T3SS) [3–5]. Note that, despite having the same name, bacterial flagella are distinct in form, function and evolution from both archaeal and eukaryotic flagella.
- biosynthesis of a molecular syringe that mediates the movement of bacterial 'effector proteins' into eukaryotic cells (a process known as 'translocation'; mediated by non-flagellar T3SSs) [1, 2, 6]. Effector proteins subvert host cell biology to the bacterium's advantage.

In both cases, the ATPase-powered secretion system is woven into a larger apparatus that includes a hollow filamentous organelle – in the case of the flagellum, the flagellar hook and filament; in case of non-flagellar systems, the needle and the translocation apparatus.

The archetypal bacterial flagellum is that of *Salmonella enterica* serovar *typhimurium* [3, 5]. This organelle consists of a basal body, set in the cell envelope, and two axial structures, the hook and filament, which meet at the hook-filament junction. Export of the components of the axial structures occurs through a central channel and depends on the flagellar type III secretion system, which lies in the central pore of the basal body MS ring and utilizes the energy of ATP hydrolysis by a peripheral hexameric ATPase, FliI.

Rotation of the flagellum is powered by the proton-motive force, with the flagellar motor converting electrochemical energy into torque. The flagellar hook is a short, highly curved structure, made from 120 FlgE monomers, which functions as a universal joint, transmitting torque to the filament. The filament is a tubular structure, built from flagellin and growing to a length of around 15 microns. Rotation of the filament, a helical propeller, converts torque into thrust, conferring motility on the cell. The chemotaxis apparatus integrates diverse signals to modulate the behavior of the motor so as to propel the cell towards nutrients.

In non-flagellar T3SSs, the secretion apparatus is connected to a hollow needle that projects from the surface of the bacterium and mediates contact with host cells. The needle is linked to a peripheral translocation apparatus that creates a pore in the host plasma cell membrane. Effector proteins move through the hollow interior of the molecular syringe to be delivered directly into the host cell cytoplasm.

An Evolutionary Conundrum?

The intricate nano-engineering of type III secretion, whether deployed in the assembly of bacterial flagella or of molecular syringes, presents us with an evolutionary conundrum. How could such elaborate complexity arise purely through the processes of gradual change and natural selection, as posited by Darwin's theory of evolution? At first glance, the idea seems scarcely credible. As Darwin himself pointed out [7]: 'Nothing at first can appear more difficult to believe than that the more complex organs and instincts should have been perfected, not by means superior to, though analogous with, human reason, but by the accumulation of innumerable slight variations, each good for the individual possessor.' Building on this sense of amazement at the complexity of nature, Behe and others in the intelligent design (ID) movement have gone one step further and claimed that bacterial flagella *cannot* have evolved. Instead, they claim, these organelles *must be* the work of an intelligent designer. At the heart of their argument is the notion that bacterial flagella show 'irreducible complexity'; that is, they are so complex that they can only function when all of their components are present and therefore cannot have evolved in gradual steps from a simpler assemblage that did not contain the full machinery [8–11].

These ID arguments were discussed – and dismissed – during the recent Kitzmiller versus Dover trial in Pennsylvania [12]. They have also been addressed in print [13]. Given the constraints of space, here we can only briefly outline some of the mechanisms that have allowed complexity to emerge from simplicity and present three case studies in the evolution of bacterial flagella

and type III secretion systems. Readers are referred to recent reviews for more extensive discussion of the evolution of complexity, of bacterial flagella and of type III secretion [13–26].

From Simplicity to Complexity

From scrutiny of the structural features and sequences of proteins from bacterial flagellar systems and non-flagellar T3SSs, it is clear that several different principles or processes have enabled the complex structures we see today to develop from simpler biological entities: repetition with self-assembly, modularity and bricolage, gene duplication with subsequent diversification and recombination.

Repetition and Self-assembly

Repetition and self-assembly are the simplest concepts to grasp. At up to fifteen microns in length, the flagellar filament from *S. typhimurium* appears a remarkable structure. However, it is built up by simple repetitive incorporation of up to 30,000 subunits of a single protein, flagellin (FliC or FljB, depending on the phase of the bacterium). Similar principles apply to other supramolecular components of the flagellum or of non-flagellar T3SSs, including both hollow filamentous and ring-like structures. Furthermore, many of these structures undergo self-assembly in the absence of the rest of the system. For example, flagellin in solution can spontaneously self-assemble into filamentous structures [27] (although at least in *S. typhimurium*, the filament cap aids polymerization into the growing filament – but some systems lack a filament cap). Similarly, PrgH, a component of the needle complex in the *Salmonella* Spi1 T3SS, can undergo self-assembly into a distinctive tetrameric structure that may be an early intermediate of T3SS assembly [28]. Furthermore, PrgH and another component of needle complex, PrgK, can, in the absence of other T3SS components, oligomerize into ring-shaped structures identical in appearance and size to the base of the needle complex [28]. This ability of single molecules or pairs of proteins to self-assemble into such elaborate formations in the absence of the rest of the system clearly undermines the irreducible complexity argument. Indeed, far from being an oddity, studies on artificial proteins, particularly those with coiled coils, have shown that spontaneous polymerization into filamentous structures is a common emergent property of such proteins, that requires no explicit design [29].

Modularity

The evidence for modularity within bacterial flagellar systems stands as a powerful counter-argument against the ID concept of irreducible complexity.

Kenneth Miller has pointed out in his testimony at the Kitzmiller versus Dover trial and elsewhere that the flagellar type III secretion system constitutes a functionally discrete subsystem capable of playing a useful function (protein secretion) in the absence of the rest of the flagellar apparatus [13]. However, the chemotaxis apparatus also shows modularity in that bacterial flagellar systems can exist without it (e.g. in *Aquifex aeolicus*), but more tellingly, it is shared with archaeal flagella that otherwise show no relationship at all to bacterial flagella [30, 31].

Molecular Bricolage

Nobel laureate François Jacob used the term *bricolage* to describe the apparently cobbled-together character of many biological structures [32]. This term, from the French verb *bricoler*, to tinker or fiddle, captures the notion of a trial-and-error approach to construction that puts the objects at hand to new and unexpected uses, creating the complex and exotic from the commonplace. Bricolage has clearly been at work during the evolution of the bacterial flagellum, in that many flagellar components have been fashioned from recycled parts that occur elsewhere in nature [17]. For example, the flagellar motor components MotA/MotB are homologous to components of the Ton and Tol systems (ExbB/ExbD and TolQ/TolR) – systems that also exploit an ion-motive force, but not for motility. Instead they use it to drive active transport of substrates across the outer-membrane. Similarly, the flagellar sigma factor is evidently homologous to many other non-flagellar sigma factors. The ATPase at the heart of flagellar and non-flagellar type III secretion is unquestionably related to the catalytic subunits of the membrane-associated F-type and V-type ATPases, and to the trancriptional terminator Rho. FlgA, which plays a role in assembly of the P-ring, is homologous to CpaB, a protein involved in type IV pilus assembly, while FlgJ shares the 'amidase_4' domain with many other bacterial proteins. Many chemotaxis proteins contain response regulator domains, which are found in other bacterial proteins. The YscD family of non-flagellar T3SS proteins contains an FHA domain and BON domains, both of which are widespread elsewhere in nature.

Gene Duplication

It is well accepted that gene duplication is a route to the creation of new gene and protein functions [21–23, 26]. Creation of a second copy of a gene, leaving the original gene intact, allows the new gene to diversify, exploring new functional landscapes unhindered by the constraints of maintaining an essential role. Bacterial flagella provide several examples of gene duplication followed by diversification leading to increased functional sophistication. In many systems there is more than one flagellin. In such cases, sequence analysis provides

unequivocal evidence of common ancestry for the multiple flagellins. Furthermore, homology searches provide solid support for a similar relationship between flagellins and the FlgL/HAP3 family of hook-associated proteins. A series of proteins from the flagellar rod, hook and filament (FlgB, FlgC, FlgE, FlgF, FlgG, HAP1/FlgK) also show sequence similarities indicative of descent from a shared ancestral precursor. Thus, the axial sub-structure of the bacterial flagellum has clearly evolved by multiple rounds of gene duplication and subsequent diversification, starting with no more than two proteins (a proto-flagellin and a proto-rod/hook protein) that were presumably capable of self-assembly into a primordial rod/filament. This entirely credible evolutionary route from simplicity to complexity decisively undermines any arguments about the implausibility of gradualism in flagellar evolution. Furthermore, as there are now clear precedents for the reconstruction and characterization of primordial ancestral proteins [33], dissection of the evolutionary and functional trajectories taken by these two protein families is now clearly within the realm of laboratory experimentation.

The axial sub-structure of non-flagellar T3SSs, the rod and needle, is much simpler than that of the flagellum. However, even here it is possible to see evidence of gene duplication and diversification at work, in that the protein that makes up the shaft of the needle (PrgI, MxiH etc.) is homologous to a second protein within the same system that makes up the inner rod within the base of the needle (PrgJ, MxiI etc.) [34]. Similarly, there is evidence for gene duplication in some non-axial components of the flagellum – for example, the shared SpoA domain common to FliM and FliN.

Recombination

Flagellins provide an interesting example of the balance between conservation and variation in protein evolution [16]. All flagellins possess conserved N- and C-terminal domains that mediate the inter-subunit interactions responsible for filament assembly. Conversely, the sequence between these two domains, which encodes the surface-exposed component of the flagellar filament, is highly variable, ranging in length from effectively zero to 800 residues [16]. The origins and extent of sequence and structural variation in these central domains remain obscure. In the well-characterised flagellar systems from enteric bacteria, it is clear that these sequences confer H-antigen sub-type specificity. Furthermore, within a single species such as *E. coli*, innovation in these sequences appears to arise from recombination between existing flagellar subtypes [35]. However, the broader role of molecular cut-and-paste mechanisms in generating variable surface-exposed domains in flagellins remains unexplored. Similarly, we have little idea as to the functional constraints on the variability seen in these domains: is variation driven by the need for avoidance

of immune or phage attack or is it instead tied into variation in the mechanical requirements of flagellar filaments operating in diverse environments (e.g. high versus low viscosity)? Similar arguments apply to the recently discovered sequence variability in the equivalent domains from the EspA-like proteins [14, 36].

Case Study 1: Chickens, Eggs and Trees

The many similarities between components of flagellar and non-flagellar T3SSs, in sequence and supramolecular structure, provide compelling evidence for an evolutionary link between them. However, soon after the discovery of this link, key figures in the field of T3SS research divided into two camps in their interpretation of the 'chicken or egg' question of which class of system came first. Some claimed that non-flagellar T3SSs were likely to have been derived from flagella [37, 38], on the grounds that motility is an ancient bacterial trait while translocation of effectors requires eukaryotic hosts and therefore must be more recent. Others were of the opinion that the two kinds of system have evolved in parallel [39].

The ID movement's interest in the origins of the bacterial flagellum meant that the evolutionary relationship between the two types of T3SSs attracted additional controversy. If the two complex systems share a subset of their components because of descent from a common ancestor, then one might argue that neither can be thought of as being truly irreducible. Thus, the ID movement quickly adopted the stance that non-flagellar T3SSs must be degenerate forms of the flagellum, derived by simplification from this supposedly designed and irreducibly complex system [11]. In response, critics of ID have tended to reject the 'flagella first' concept, and have even embraced the opposite view, namely that non-flagellar T3SSs are actually precursors of bacterial flagella [13].

Molecular Phylogenies and Arguments from Cladistics

Two studies of the molecular phylogeny of protein sequences conserved in both flagellar and non-flagellar T3SSs have shed some light on this controversy [40, 41]. Both studies used similar methodologies and arrived at similar tree topologies (fig. 2). However, surprisingly, there were differences of opinion in the interpretation of these data. We have argued consistently that these phylogenetic studies demonstrate that each class of system is monophyletic, and therefore each must have evolved in parallel from a common ancestor, probably one that was simpler than any extant system. Others have interpreted the same data to support the idea that non-flagellar T3SSs are descended from bacterial flagella [40], and have then posited a false dichotomy whereby rejection of the

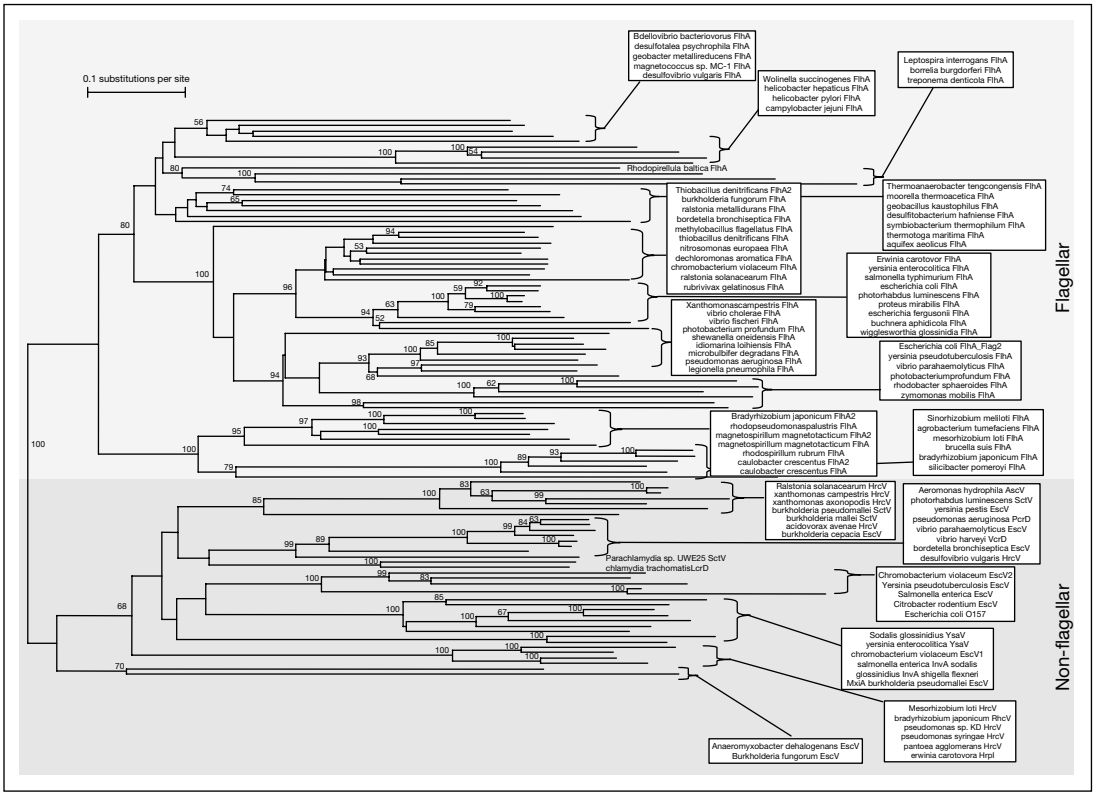


Fig. 2. Phylogenetic tree of FlhA/EscV gene homologs in type III secretion systems. Alignment was performed by ClustalW and hand edited to remove ambiguous positions. Tree was generated using the neighbor-joining algorithm correcting for multiple substitutions and ignoring positions with gaps. Numbers showing bootstrap support values exceeding 50% are presented. The *Chlamydia* FlhC sequence was omitted from the analysis because it lies on a very long branch and creates neighbor-joining artifacts.

'flagella-first' argument implies support for the notion that flagella are derived from non-flagellar T3SSs [42].

Although traditional molecular phylogenetic analyses, based on multiple alignments of homologous protein sequences from multiple systems, provide one route to the reconstruction of the evolutionary history of type III secretion systems, we have argued that an alternative, cladistic approach, based on the analysis of higher-order features (gene/protein content and arrangement, with consequent effects on T3SS structure and function) might also be informative [14]. It is clear that this approach, twinned with arguments from parsimony, lends support to the idea that the flagellar and non-flagellar classes are sister

groups that evolved divergently from a common ancestor, in that both classes share distinctive derived characters (synapomorphies) common to all members of the class. In particular, non-flagellar systems possess a translocation apparatus and components associated with the inner and outer membranes (the FHA domain in the YscD-like proteins, the secretin domain in the YscC-like proteins) that are absent from the flagellar systems, while flagellar systems share the flagellar motor, filament and sigma factor that are absent from the non-flagellar systems.

'Difficulties on Theory': EspA and FliK

As noted, molecular phylogenetic studies and cladistics support the notion of divergent evolution of flagellar and non-flagellar systems from a common ancestor, with each of the two major classes exhibiting the property of monophyly [14]. However, the existence of homologous components that are shared between *all* flagellar systems but only *some* sub-classes of non-flagellar systems presents problems.

The EspA filament was first discovered in the Esc-Esp T3SS encoded by the locus for enterocyte effacement in selected strains of *E. coli* and *Citrobacter rodentium* [43]. Since its discovery, evidence has steadily accumulated to suggest that EspA and the EspA filament are homologous to flagellin and the flagellar filament respectively. At the time of writing, the evidence is strong, but not absolutely conclusive. Final proof (or disproof) will arrive only when a three-dimensional structure of EspA within the filament becomes available.

The EspA protein from *E. coli*, appeared, at first, to be one of a kind. However, now thanks to genome sequencing, we know that EspA-like proteins are a characteristic of all members of the Esc-Ssa sub-group of non-flagellar T3SSs, occurring in organisms as diverse as *Salmonella enterica, Edwardsiella ictulari, Shewanella baltica, Chromobacterium violaceum, Yersinia frederiksenii, Yersinia bercovieri,* and *Sodalis glossinidius*. However, one looks in vain for any inkling of an EspA homologue in any of the four or more other sub-classes of non-flagellar T3SS. Thus, assuming that EspA really is a flagellar homologue, one is left with three plausible scenarios:

1. The presence of an EspA/FliC-like protein/filament is the ancestral state for all T3SSs; the Esc-Ssa sub-group represents the earliest diverging branch of the non-flagellar systems; and the EspA/FliC-like protein was subsequently lost in the branch leading to all other NF-T3SSs. This scenario should not be seen as supporting the 'flagellum-first' position, as the ancestral EspA/FliC filament need not have played any role in motility. Instead, in this scenario, the EspA filament provides a model for how an ancestral filament might have functioned for purposes other than locomotion (e.g. adhesion and/or targeted protein secretion). The major problem

with this scenario is the counter-intuitive idea that something as interesting as the EspA filament could have been lost in the evolution of most non-flagellar systems.

2. The ancestral system lacked any FliC/EspA protein. Instead, this protein arose after the divergence of the common ancestor of the flagellar systems and the Esc-Ssa sub-group from the ancestor of the remaining non-flagellar systems. However, this scenario is not consistent with the monophyletic properties of the flagellar and non-flagellar systems, as revealed by traditional molecular phylogenetic studies.
3. Flagellin evolved in the flagellar systems after their divergence from the non-flagellar systems, and EspA was recruited into a sub-set of non-flagellar systems by horizontal gene transfer. It is hard to distinguish this scenario from scenario 1, followed by recombination between a non-flagellar EspA-like protein and a proteobacterial flagellin.

FliK is a flagellar protein involved in hook length control. YscP is a protein from the Ysc-Yop non-flagellar T3SS that is involved in needle-length control. Whether equivalent mechanisms govern the mechanism of action of these proteins in flagellar and non-flagellar systems has been the source of some controversy [44, 45]. Nonetheless, sequence analyses clearly show that they share homologous C-terminal domains [15, 46]. The problem is that proteins bearing this C-terminal domain have been identified only in close relatives of the Ysc-Yop system. Proteins which appear to lack this domain, but which might play equivalent roles have been identified in some systems in the Inv-Spa sub-group of non-flagellar T3SSs [14], but no plausible equivalents have been identified in the other sub-classes of non-flagellar systems (the Hrp1, Hrp2, Esc-Ssa, chlamydial sub-classes etc).

Thus, as with FliC/EspA-like proteins, coming up with an evolutionary model that reconciles the distribution of FliK/YscP-like proteins in type III secretion systems and pattern of evolution deduced from the molecular phylogenetic data is not easy. As it is likely that the hook and needle are homologous structures, it seems plausible that they would share an ancestral length-determining mechanism. However, if FliK/YscP-like proteins are an ancestral feature of type III secretion, their presence in the Ysc family but their apparent loss in the lineage leading to the Esc-Ssa and other families (Hrp1, Hrp2, chlamydia etc.) implies that the Ysc family diverged from all other non-flagellar T3SSs earlier than the Esc-Ssa family, which stands in contradiction to scenario 1 for the evolution of FliC/EspA proteins.

These apparent inconsistencies in our understanding of the evolution of type III secretion, especially the fruitless 'chicken or the egg' argument as whether flagellar or non-flagellar systems came first, will soon, we hope, be resolved, as we gain additional structural and functional insights to inform our

assignments of homology and ever more examples of type III secretion systems become available, including systems that combine features of the flagellar and non-flagellar classes and/or that are substantially simpler than currently known examples (see discussion of myxococcal systems below).

Case Study 2: New Homologies?

One of the claims put forward in support of the notion of ID is that 'the other thirty proteins in the flagellar motor (that are not present in the type III secretion system) are unique to the motor and are not found in any other living system' [11]. We have already largely addressed this claim above in our discussions of the role of gene duplication and bricolage in flagellar evolution. However, this 'God-of-the-gaps' assertion is also a hostage to fortune, in that as additional information comes in from bioinformatics and laboratory-based studies, any remaining orphan proteins are likely to become the subjects of new assignments of homology. Here, we present evidence for two new tentative claims of previously unrecognized homology that await experimental confirmation.

Is FliJ Homologue of YscO?

FliJ is a small soluble component of the *Salmonella* flagellar type III secretion protein of unknown function [47–49]. It is encoded by *fliJ*, which sits between the gene for the T3SS ATPase, *fliI*, and the gene for the hook-length control protein, *fliK*. In the Ysc-Yop system, the genes encoding the homologues of *fliI* and *fliK* are *yscN* and *yscP*. As with their flagellar equivalents, these two genes sit either side of a gene of unknown function, in this case *yscO*. The associated protein YscO is similar in size to FliJ (154 amino-acids versus 147) and is known to play a role in type III secretion [50]. Both FliJ and YscO are predicted to be entirely alpha-helical and to contain extensive regions of coiled coil. Unfortunately, these properties mean that homology searches with either protein quickly attract many other coiled coil or alpha-helical proteins, so that it is impossible using BLAST or PSI-BLAST to achieve statistically significant evidence for an assignment of homology. However, further evidence for a conserved protein family comes from consideration of gene order in other non-flagellar T3SSs, where a gene encoding a protein with similar properties is located downstream of the ATPase gene in most if not all cases. A multiple alignment of such proteins also suggests that they might indeed be members of a common protein family (fig. 3).

Even if sequence analyses cannot conclusively answer the question of whether FliJ and YscO are homologues, we can make experimentally testable predictions on the basis of this hypothesis. The C-terminal region of FliJ is

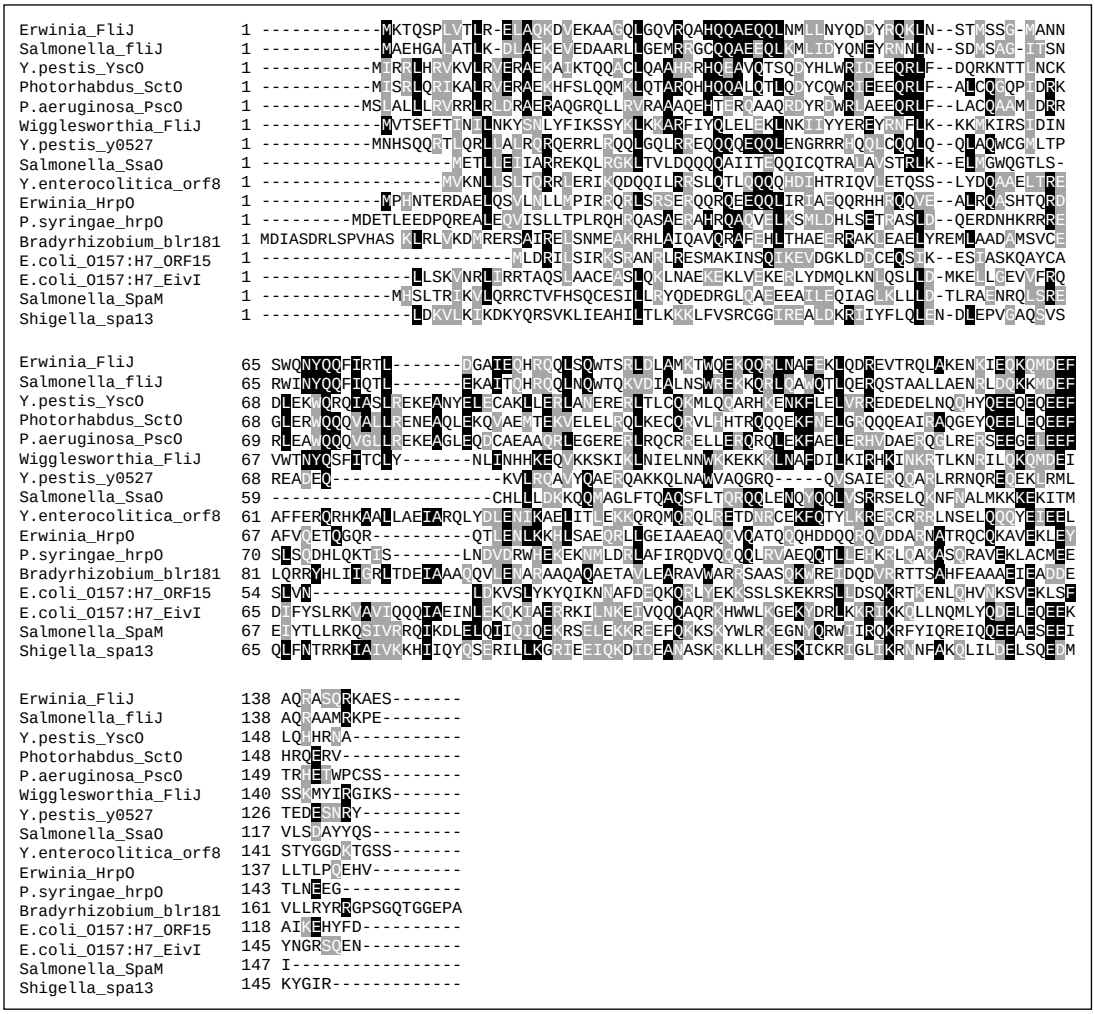


Fig. 3. ClustalW multiple sequence alignment of FliJ/YscO homologs.

known to interact with the N-terminal region of FliH, a negative regulator of the ATPase FliI [51]. FliJ also interacts with the cytoplasmic domain of FlhA and with FliM [47, 49]. If YscO is indeed a homologue of FliJ, then we would predict that it should partake in similar protein-protein interactions with homologous proteins in the Ysc-Yop system. In particular, the C-terminus of YscO should interact with the N-terminus of YscL (the FliH homologue) and YscO

should also interact with LcrD/YscV and YscQ. Furthermore, once the structures and functions of both proteins have been elucidated, we can predict that profound similarities will be found, indicative a common ancestry. Alternatively, if our hypothesis is wrong, we would expect to see none of these interactions and only superficial functional or structural similarities.

Origins of the Flagellar Motor: From Ion Channel to Rotor?

The flagellar motor converts electrochemical energy into torque through an interaction between two components: the stator and the rotor. The stator consists of multiple copies of two proteins, MotA and MotB, which assemble into a structure associated with the inner membrane and anchored to peptidoglycan, so that it remains stationary. The rotor consists of multiple copies of FliG, which together with FliM and FliN form the C ring, mounted on the cytoplasmic face of the MS ring.

The MotA/MotB and FliM/FliN proteins have obvious homologues outside the flagellum [17], but FliG remains an apparent orphan, even though a partial structure is available [52]. However, a search for homologues of the sequence corresponding to the known structure using the highly sensitive, iterative PSI-BLAST facility at the National Center for Biotechnology Information is revealing. When the search is carried out under default conditions (aside from setting the number of alignments and descriptions reported to 1000), the first couple of iterations return lots of FliG proteins, plus an unusual protein from *Chlamydia* that consists of a FliG-like fragment stuck on to FliF. However, later iterations report significant similarity between FliG and multiple homologues of a bacterial magnesium transport protein, MgtE, particularly the intracellular N-terminal domain of that protein (PFAM domain PF03448).

The MgtE family was discovered by Maguire and colleagues [53]. These ca. 40-kDa proteins occur in a range of bacteria and consist of an intracellular N-terminal domain and five to six transmembrane segments [54]. Although they are known to translocate Mg^{2+} and Co^{2+} into the cell, the primary function of MgtE family in bacteria is not yet fully understood. However, the possibility of an evolutionary link between a membrane-associated transporter of divalent cations and a flagellar protein involved in generating torque from H^+ or Na^+ transport is intriguing.

To be convinced that the relationship is real, it would be helpful if we could show that the search works in reverse, i.e. that one can get from an MgtE protein to FliG. Using the N-terminal domain from the *Y. pestis* MgtE protein as starting point for a fresh PSI-BLAST search under the same conditions, after half a dozen of iterations several FliG proteins are reported with scores above the threshold for inclusion in subsequent iterations and within a few more iterations, the FliG sequence of known structure has crossed the threshold.

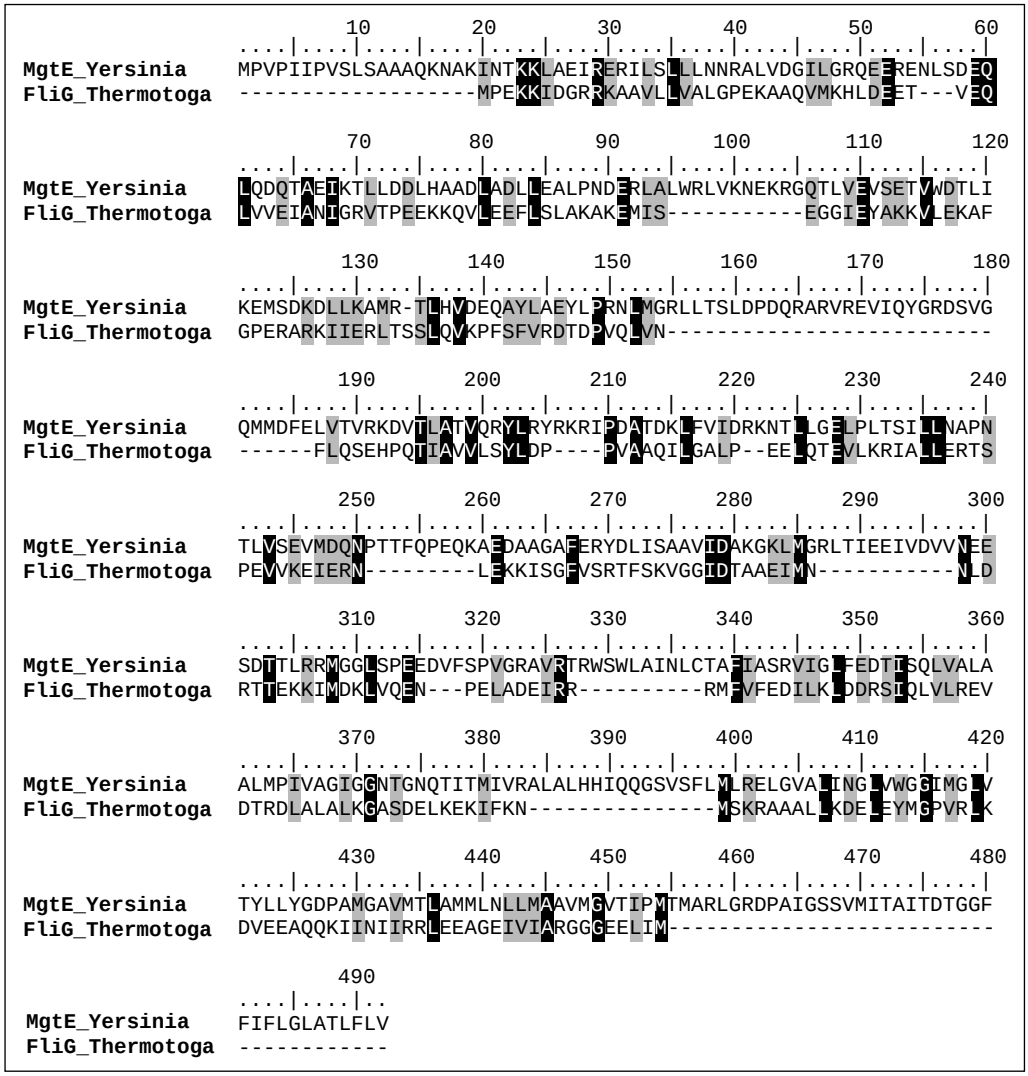


Fig. 4. ClustalW pairwise alignment of MgtE from *Yersinia pestis* and FliG from *Thermotoga maritime*.

Additional support for homology between FliG and the N-terminal domain of MgtE comes from searches of domain databases that use approaches other than PSI-BLAST. Searches of the PFAM and CDD databases with the FliG sequence report hits to the MgtE domain, albeit with unimpressive significance scores.

Given the very low levels of sequence identity reported by PSI-BLAST between FliG and MgtE (<20%), and the difficulties in obtaining a plausible multiple alignment between sequences from the two protein families, it is probably premature to draw a firm conclusion that they are indeed homologues. However, these findings do provide a solid justification for solving the structure of the MgtE N-terminal domain and for further structural studies on FliG.

Case Study 3: The Search for Outliers

In 2002, Foultier et al. [55] concluded from molecular phylogenetic analyses that the proteobacterial NF-T3SSs could be classified into five large classes: the Ysc group, the Esc/Ssa group, the Hrp1 and Hrp2 groups, and the Inv/Mxi/Spa groups. In addition, they found two orphan systems that did not comfortably fit into any of these large classes, the *Bordetella* T3SS (which in some analyses fell within the Ysc group) and the rhizobial system. Largely similar conclusions have been reached in other such analyses [40, 41, 56].

Chlamydial T3S genes have attracted attention for three reasons [57]: (1) rather than one gene cluster encoding the entire secretion system, there are several; (2) the T3SS genes do not deviate from the chromosomal average in G+C content, suggesting that they originated here, or have been resident in these genomes for a very long time; (3) there are several pairs of paralogues, where one T3S component looks as if it belongs to a non-flagellar system, while the other is more like equivalent flagellar proteins.

More recently, two genomes from the myxococci harboring T3SS genes have become available, one from *Myxococcus xanthus* the other from *Anaeromyxobacter dehalogenans*. These are of interest because, like the chlamydial systems, these myxococcal systems show some unusual features. In our phylogenetic analysis of T3SSs, one of the systems from *Anaeromyxobacter dehalogenans* appears as an outlier. Furthermore, in both myxococcal genomes there are two T3SS gene clusters, one apparently complete, the other lacking key components such as the ATPase. Blast searches with these proteins in some cases report non-flagellar T3SS proteins as the highest-scoring hits, while in other cases, they report the highest similarity to flagellar proteins. In *Myxococcus xanthus*, there are no genes for components of the flagellar hook or filament, and an apparently flagellar-like cluster encodes proteins with TPR and FHA domains which are otherwise restricted to non-flagellar systems. One cannot rule out from sequence analyses alone the possibility that these gene clusters might represent degenerate non-functional remnants of larger loci. However, given their potential roles as ‘missing links’ in the evolution of type III secretion, determining whether these loci still encode functional secretion

systems, and what role such systems play in their host cell physiology must surely rate as a high priority for myxococcal research in the post-genomic era.

Conclusions

Type III secretion systems provide excellent case studies in the evolution of complexity from simplicity. However, detailed analysis of the evolution of the sequences, gene and protein repertoires and macromolecular structures associated with these systems has scarcely begun. One key challenge will be determining the polarity of changes – that is whether the presence of a feature in one system but not another represents loss of an ancestral feature or gain of a novel character. Fortunately, the steady accumulation of sequence data from genome sequencing promises to provide us with the raw material for new insightful analyses. Finally, we conclude that bacterial flagella and type III secretion systems provide no challenge whatsoever to the Darwinian paradigm. As Darwin himself pointed out [7]: 'If it could be demonstrated that any complex organ existed, which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down. But I can find out no such case.' And nor can we!

Note Added in Proof

The structure of the soluble part of MgtE from *Enterococcus faecalis* has now been solved (PDB: 2OUX) and reveals that the N-terminal part of the protein shares a similar fold with FliG, supporting the assignment of homology deduced from sequence analyses.

References

1 Journet L, Hughes KT, Cornelis GR: Type III secretion: a secretory pathway serving both motility and virulence (review). Mol Membr Biol 2005;22:41–50.
2 Mota LJ, Cornelis GR: The bacterial injection kit: type III secretion systems. Ann Med 2005;37:234–249.
3 Macnab RM: Type III flagellar protein export and flagellar assembly. Biochim Biophys Acta 2004;1694:207–217.
4 Minamino T, Namba K: Self-assembly and type III protein export of the bacterial flagellum. J Mol Microbiol Biotechnol 2004;7:5–17.
5 Macnab RM: How bacteria assemble flagella. Annu Rev Microbiol 2003;57:77–100.
6 Tobe T, et al.: An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. Proc Natl Acad Sci USA 2006;103:14941–14946.
7 Darwin C: The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, ed 1. John Murray, London, 1859.
8 Behe M: The challenge of irreducible complexity. Nat Hist 2002;111:74.

9 Behe MJ: Darwin's Black Box: The Biochemical Challenge to Evolution. Free Press, New York, 1996.
10 Minnich SA: Expert witness report in Kitzmiller v. Dover. National Center for Science Education, 2005.
11 Minnich SA, Meyer SC: Genetic analysis of coordinate flagellar and type III regulatory circuits in pathogenic bacteria; in Collins MW, Brebbia CA (eds): Design and Nature II: Comparing Design in Nature with Science and Engineering. Wessex Institute of Technology Press, Southampton/Boston, 2004.
12 Jones JEI: Kitzmiller et al. versus Dover Area School District et al Memorandum Opinion, in United States District Court, Middle District of Pennsylvania. Case No. 04cv2688 (Harrisburg, PA 2005).
13 Miller KR: The flagellum unspun: the collapse of "irreducible complexity"; in Dembski W, Ruse M (eds): Debating Design: From Darwin to DNA. Cambridge University Press, New York, 2004.
14 Pallen MJ, Beatson SA, Bailey CM: Bioinformatics, genomics and evolution of non-flagellar type-III secretion systems: a Darwinian perspective. FEMS Microbiol Rev 2005;29:201–229.
15 Pallen MJ, Penn CW, Chaudhuri RR: Bacterial flagellar diversity in the post-genomic era. Trends Microbiol 2005;13:143–149.
16 Beatson SA, Minamino T, Pallen MJ: Variation in bacterial flagellins: from sequence to structure. Trends Microbiol 2006;14:151–155.
17 Pallen MJ, Matzke NJ: From the origin of species to the origin of bacterial flagella. Nat Rev Microbiol 2006;4:784–790.
18 Dawkins R: Climbing mount improbable, 1st American ed. Norton, New York 1996.
19 Dawkins R: The God Delusion. Houghton Mifflin Co, Boston, 2006.
20 Thornhill RH, Ussery DW: A classification of possible routes of Darwinian evolution. J Theor Biol 2000;203:111–116.
21 Ganfornina MD, Sanchez D: Generation of evolutionary novelty by functional shift. Bioessays 1999;21:432–439.
22 True JR, Carroll SB: Gene co-option in physiological and morphological evolution. Annu Rev Cell Dev Biol 2002;18:53–80.
23 Orengo CA, Thornton JM: Protein families and their evolution – a structural perspective. Annu Rev Biochem 2005;74:867–900.
24 Rison SC, Thornton JM: Pathway evolution, structurally speaking. Curr Opin Struct Biol 2002;12:374–382.
25 Adami C, Ofria C, Collier TC: Evolution of biological complexity. Proc Natl Acad Sci USA 2000;97:4463–4468.
26 Hancock JM: Gene factories, microfunctionalization and the evolution of gene families. Trends Genet 2005;21:591–595.
27 Auvray F, Thomas J, Fraser GM, Hughes C: Flagellin polymerisation control by a cytosolic export chaperone. J Mol Biol 2001;308:221–229.
28 Kimbrough TG, Miller SI: Contribution of *Salmonella typhimurium* type III secretion components to needle complex formation. Proc Natl Acad Sci USA 2000;97:11008–11013.
29 Woolfson DN: The design of coiled-coil structures and assemblies. Adv Protein Chem 2005;70:79–112.
30 Faguy DM, Jarrell KF: A twisted tale: the origin and evolution of motility and chemotaxis in prokaryotes. Microbiology 1999;145(Pt 2):279–281.
31 Szurmant H, Ordal GW: Diversity in chemotaxis mechanisms among the bacteria and archaea. Microbiol Mol Biol Rev 2004;68:301–319.
32 Jacob F: Evolution and tinkering. Science 1977;196:1161–1166.
33 Cai W, Pei J, Grishin NV: Reconstruction of ancestral protein sequences and its applications. BMC Evol Biol 2004;4:33.
34 Pallen MJ, Beatson SA, Bailey CM: Bioinformatics analysis of the locus for enterocyte effacement provides novel insights into type-III secretion. BMC Microbiol 2005;5:9.
35 Wang L, Rothemund D, Curd H, Reeves PR: Species-wide variation in the *Escherichia coli* flagellin (H-antigen) gene. J Bacteriol 2003;185:2936–2943.
36 Betts HJ, Chaudhuri RR, Pallen MJ: An analysis of type-III secretion gene clusters in *Chromobacterium violaceum*. Trends Microbiol 2004;12:476–482.

37 Galan JE, Collmer A: Type III secretion machines: bacterial devices for protein delivery into host cells. Science 1999;284:1322–1328.
38 Macnab RM: The bacterial flagellum: reversible rotary propellor and type III export apparatus. J Bacteriol 1999;181:7149–7153.
39 Aizawa SI: Bacterial flagella and type III secretion systems. FEMS Microbiol Lett 2001;202:157–164.
40 Nguyen L, Paulsen IT, Tchieu J, Hueck CJ, Saier MH Jr.: Phylogenetic analyses of the constituents of type III protein secretion systems. J Mol Microbiol Biotechnol 2000;2:125–144.
41 Gophna U, Ron EZ, Graur D: Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. Gene 2003;312:151–163.
42 Saier MH Jr.: Evolution of bacterial type III protein secretion systems. Trends Microbiol 2004;12:113–115.
43 Knutton S, Rosenshine I, Pallen MJ, Nisan I, Neves BC, et al: A novel EspA-associated surface organelle of enteropathogenic *Escherichia coli* involved in protein translocation into epithelial cells. EMBO J 1998;17:2166–2176.
44 Journet L, Agrain C, Broz P, Cornelis GR: The needle length of bacterial injectisomes is determined by a molecular ruler. Science 2003;302:1757–1760.
45 Makishima S, Komoriya K, Yamaguchi S, Aizawa SI: Length of the flagellar hook and the capacity of the type III export apparatus. Science 2001;291:2411–2413.
46 Agrain C, Callebaut I, Journet L, Sorg I, Paroz C, et al: Characterization of a Type III secretion substrate specificity switch (T3S4) domain in YscP from *Yersinia enterocolitica*. Mol Microbiol 2005;56:54–67.
47 Fraser GM, Gonzalez-Pedrajo B, Tame JR, Macnab RM: Interactions of FliJ with the *Salmonella* type III flagellar export apparatus. J Bacteriol 2003;185:5546–5554.
48 Minamino T, Chu R, Yamaguchi S, Macnab RM: Role of FliJ in flagellar protein export in *Salmonella*. J Bacteriol 2000;182:4207–4215.
49 Gonzalez-Pedrajo B, Minamino T, Kihara M, Namba K: Interactions between C ring proteins and export apparatus components: a possible mechanism for facilitating type III protein export. Mol Microbiol 2006;60:984–998.
50 Payne PL, Straley SC: YscO of *Yersinia pestis* is a mobile core component of the Yop secretion system. J Bacteriol 1998;180:3882–3890.
51 Gonzalez-Pedrajo B, Fraser GM, Minamino T, Macnab RM: Molecular dissection of *Salmonella* FliH, a regulator of the ATPase FliI and the type III flagellar protein export pathway. Mol Microbiol 2002;45:967–982.
52 Brown PN, Hill CP, Blair DF: Crystal structure of the middle and C-terminal domains of the flagellar rotor protein FliG. EMBO J 2002;21:3225–3234.
53 Smith RL, Thompson LJ, Maguire ME: Cloning and characterization of MgtE, a putative new class of Mg^{2+} transporter from *Bacillus firmus* OF4. J Bacteriol 1995;177:1233–1238.
54 Maguire ME: Magnesium transporters: properties, regulation and structure. Front Biosci 2006;11:3149–3163.
55 Foultier B, Troisfontaines P, Muller S, Opperdoes FR, Cornelis GR: Characterization of the ysa pathogenicity locus in the chromosome of *Yersinia enterocolitica* and phylogeny analysis of type III secretion systems. J Mol Evol 2002;55:37–51.
56 He SY, Nomura K, Whittam TS: Type III protein secretion mechanism in mammalian and plant pathogens. Biochim Biophys Acta 2004;1694:181–206.
57 Kim JF: Revisiting the chlamydial type III protein secretion system: clues to the origin of type III protein secretion. Trends Genet 2001;17:65–69.

Mark J. Pallen
University of Birmingham Medical School
Birmingham, B15 2TT
United Kingdom
Tel. +44 121 414 7163, Fax +44 121 414 3454, E-Mail m.pallen@pallen@bham.ac.uk

Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 48–65

Comparative Genomics and Evolutionary Trajectories of Viral ATP Dependent DNA-Packaging Systems

A.M. Burroughs[a,b], *L.M. Iyer*[a], *L. Aravind*[a]

[a]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, [b]Bioinformatics Program, Boston University, Boston, MA, USA

Abstract

We present an overview of comparative genomics of ATP-dependent DNA packaging systems of viruses. Several distinct ATPase motors and accessory proteins have been identified in DNA-packaging systems of viruses such as terminase-portal systems, the φ29-like packaging apparatus, and packaging systems of lipid inner-membrane-containing viruses. Sequence and structure analysis of these proteins suggest that there were two major independent innovations of ATP-dependent DNA packaging systems in the viral universe. The first of these utilizes a HerA/FtsK superfamily ATPase and is seen in prokaryotic viruses with inner lipid membranes, large eukaryotic nucleo-cytoplasmic DNA viruses (including poxviruses) and a group of eukaryotic mobile DNA transposons. We show that ATPases of the φ29-like packaging system are also divergent versions of the HerA/FtsK superfamily that functions in viruses without an inner membrane. The second system, the terminase-portal system, is dominant in prokaryotic tailed viruses and typically functions with linear chromosomes. The large subunit of this system contains a distinct ATPase domain and a C-terminal nuclease domain of the RNAse H fold. We discuss the classification of these ATPases within the P-loop NTPases, genomic demography and positioning of their genes in the viral chromosome. We show that diverse portal proteins utilized by these systems share a common evolutionary origin and might have frequently displaced each other in evolution. Examination of conserved gene neighborhoods indicates repeated acquisition of Helix-turn-Helix domain-containing terminase small subunits and a third accessory component, the MuF protein. Adenoviruses appear to have evolved a third packaging ATPase, unique to their lineage. Relationship between one major type of packaging ATPases and cellular chromosome pumps like FtsK suggests an ancient common origin for viral packaging and cellular chromosome partitioning systems.

Copyright © 2007 S. Karger AG, Basel

Proper segregation of chromosomes and their partitioning into daughter cells or capsids is a common problem faced by cellular and viral replicons. While diverse solutions to this problem have evolved in different viruses, they may all be categorized under two broad mechanistic themes [1, 2]. Most RNA viruses and several small DNA viruses do not appear to require an active energy-dependent process for packaging their genomes, and the process simply proceeds via coating of nucleic acids by capsid subunits. Coating is usually initiated by packaging signals in the form of sequence or structural features in the nucleic acid, resulting in condensation of the capsid proteins on the nucleic acid scaffold [3, 4]. In the second theme, an active ATP-dependent process loads the genome of larger double stranded (ds) DNA viruses and single stranded (ss) DNA viruses of the Inovirus family into empty capsids [2, 5, 6].

Extreme sequence divergence of viral proteins has hampered understanding of relationships between components of chromosome segregation and packaging systems of different viruses. However, recent availability of a wealth of crystal structures and complete sequences of numerous viral genomes allows us to address this problem using a variety of sequence and structure analysis techniques and comparative genomics. Some recent developments in this regard include structural studies on viral coat proteins revealing that the principal capsid or coat protein of several characterized viruses contains a distinctive β-strand fold with a β-jelly-roll topology [7, 8]. Remarkably, this structural conservation of capsid proteins transcends the diversity of viruses, which might be otherwise unrelated in terms of their genomic nucleic acid or replication and packaging mechanisms. This raised the intriguing possibility that principal capsid proteins of a notable subset of viruses might have descended from a common ancestor [8].

Similarly, sensitive sequence comparisons showed that packaging ATPase motors of diverse large eukaryotic and prokaryotic DNA viruses belong to the HerA-FtsK superfamily, which includes the DNA pumps involved in prokaryotic cellular chromosome segregation and related DNA pumps of several conjugative plasmids and transposons [5]. Packaging ATPases of the HerA/FtsK superfamily are encountered in the recently unified Nucleo-Cytoplasmic Large DNA Virus (NCLDV) assemblage and in several dsDNA phages like PRD1 and the Inovirus family [5, 6]. The other major functionally characterized ATP-dependent DNA-packaging system is the terminase-portal protein system first noticed in caudoviruses (tailed prokaryotic viruses) and herpesviruses [2, 9]. In its most basic form the system consists of the two-subunit terminase complex and a multimeric portal protein (PP) providing a conduit for nucleic acid entry into capsids. The terminase large subunit (TLS) has both ATPase activity that powers DNA translocation and nuclease activity, which cleaves the replicating

DNA into genome-sized fragments [10–13]. In addition to these systems, there are smaller families of packaging ATPases in φ29-like bacteriophages and the adenoviruses whose evolutionary affinities were previously unclear [14, 15].

In this article we build upon these previous studies to provide a synthetic overview of the protein components of various ATP-dependent phage DNA-packaging systems. We establish the evolutionary affinities of several poorly understood components and also describe new potential components. Relationships and structural features of packaging proteins presented here also throw light on various functional aspects of DNA packaging, with general implications for the origins of chromosome segregation.

Results and Discussion

The Demography of Packaging ATPases in Large DNA Viruses

Amongst large eukaryotic DNA viruses, all NCLDVs, namely poxviruses, iridoviruses, African Swine Fever Virus, phycodnaviruses and the mimivirus, share a packaging ATPase of the HerA/FtsK superfamily [5,16]. The herpesviruses contain a terminase-portal packaging system similar to the bacteriophages [9]. Additionally, a HerA/FtsK-type ATPase is also encoded by a novel DNA transposon that is widespread in *Trichomonas*, ciliate and nematode genomes [5]. The transposon also has some relationship with adenoviruses in terms of its DNA polymerase and processing protease, suggesting that it might assemble into virus like particles aided by this ATPase [5, 17]. The predicted packaging ATPase of the adenoviruses has thus far not been seen in any other viral lineage [15, 18]. Packaging ATPases, if any, of certain large DNA viruses like baculoviruses and the shrimp white spot syndrome virus are unknown, but they are unlikely to define large new lineages of packaging enzymes.

Our systematic survey of phage packaging system components in completed genomes of 288 prokaryotic DNA viruses showed that they encompass a comparable diversity in terms of their ATPases. Up to a genome size of about 20 kb there is a steady increase in the fraction of phages encoding packaging ATPases (fig. 1a). Beyond this size, 95% of phages encode a packaging ATPase. The majority of small phages lacking packaging ATPases are microviruses, which initiate their packaging through a passive interaction with a small genomically encoded polypeptide [4]. This suggests that 20 kb is the approximate size threshold above which packaging appears to require an active energy-dependent process. The most common packaging ATPase in currently available phages is the terminase-type ATPase (seen in ~70% of the phages), whereas ~15% of phages utilize a version of the HerA/FtsK ATPase superfamily (see supplementary material: SM). The presence of a terminase-type ATPase

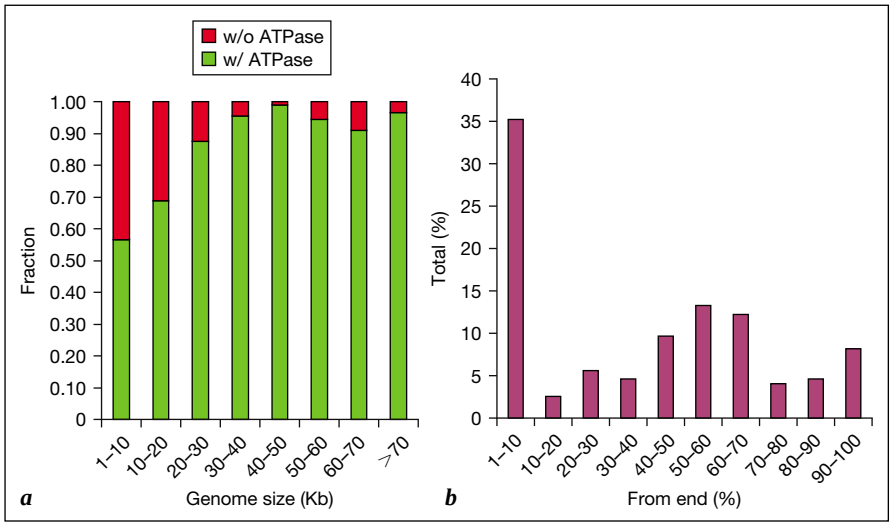


Fig. 1. Packaging ATPase presence/absence and positional distribution in viral genomes. ***a*** Presence/absence of a packaging ATPase in completely-sequenced viral genomes is depicted as a stacked column graph. Percentages of genomes containing a packaging ATPase within a certain genome size range are green columns while percentages lacking an ATPase are red. ***b*** Genome position frequency distribution of packaging ATPases from completely-sequenced viral genomes with linear chromosomes is shown as bars graph. Statistically significant preference for placement in the middle or termini of viral genomes is observed (χ^2: $p < 10^{-5}$).

is strongly correlated with the tailed capsid morphology typical of caudoviruses, the most common type of bacteriophage. The HerA/FtsK family appears to exclusively occur in phages with internal lipid membranes, such as tectiviruses, corticoviruses and *Sulfolobus* turreted virus, irrespective of their outer protein coat morphology (SM) [5, 6, 19]. These phages also often contain terminal inverted repeats in their genomes. Furthermore, 85% of viruses with terminase-type ATPases have linear chromosomes, while 68% of those with HerA/FtsK type ATPases have circular chromosomes (SM). This suggests that while each system can handle either chromosome type, there might be a preferred type for each of them.

A study of the positional distribution of genes for packaging ATPases in phages with linear genomes revealed that in 70% of the cases they are either located at an end or close to the center of the genome (fig. 1b). This unusual distribution is highly significant ($p < 10^{-5}$ by Chi-test) and appears to be related to the time of transcription of the packaging ATPase in the virus life

cycle. Placement of these genes towards the chromosome termini or in the middle may allow late transcription, thereby making the packaging apparatus available only at the last phase of the viral cycle. This bias in chromosomal position of the gene for packaging ATPases provides a contextual means of predicting potential packaging ATPases of uncharacterized viruses. We observed two archaeal globuloviruses (*Thermoproteus tenax* spherical virus 1: TTSV and *Pyrobaculum* spherical virus: PSV) with genome sizes greater than 20 kb lacking any known packaging ATPases. However, both viruses encode an uncharacterized P-loop NTPase at the termini of their genomes (TTSV: ORF1 and PSV: ORF582). Sequence searches with these proteins showed no close relation to any other ATPases involved in replication such as helicases or clamp loaders; supporting its possible role as a packaging ATPase.

Multiple Origins for Different Packaging ATPases within the P-loop NTPase Fold

All known and predicted packaging ATPases of DNA viruses belong to the P-loop NTPase fold, one of the most prevalent protein folds in both cellular and viral genomes [2, 5, 15]. Members of the P-loop NTPase fold are unified by the conserved nucleotide binding (Walker A) and Mg^{2+} binding motifs (Walker B) and belong to one of two major divisions; the KG division which includes P-loop kinases and GTPases, and the ASCE (additional strand conserved E (glutamate)) division [20, 21]. The latter division is characterized by an additional conserved acidic residue (typically a glutamate occurring immediately after the conserved Walker B aspartate) and a conserved polar residue (Sensor 1) occurring at the end of the 4th core strand of the domain [21, 22]. Examination of all characterized packaging ATPase domains, namely the TLS N-terminal domain, the HerA/FtsK ATPase domain, the φ29-like phage ATPase domain, and the putative adenoviral packaging ATPase domain revealed hallmark features of the ASCE division, indicating derivation from within this radiation of the P-loop fold [5, 15]. This observation is consistent with the fact that the majority of highly active ATPases mediating energy dependent processes in biological systems belong to the ASCE division [20–22].

However, relationships between different viral packaging ATPases and affinities to other major classes of ATPases of the ASCE group have remained largely unclear. Previous systematic analysis of the HerA/FtsK superfamily revealed that viral packaging ATPases of this superfamily do not form an exclusive virus-specific clade but are successive out-groups of the crown-group formed by cellular HerA and FtsK families [5]. The basal-most clade was comprised of packaging ATPases of filamentous inoviruses with ssDNA genomes while those from remaining diverse groups of lipid membrane-containing dsDNA viruses of prokaryotes and eukaryotes formed a large assemblage, an

immediate sister group of the cellular and plasmid members of this superfamily [5]. Preliminary sequence searches with φ29-like ATPases recovered only cognate ATPases of other related viruses and secondary structure prediction revealed that φ29-like ATPases contained an α-β unit C-terminal to strand-2 as seen in the FtsK, RecA, helicase and PilT assemblage within the ASCE division. Furthermore, we noted that φ29-like ATPases bore a conserved arginine at the base of strand-4, equivalent to identically positioned arginine fingers in HerA/FtsK ATPases. φ29-like ATPases also possessed a conserved asparagine at the end of the sensor-1 strand, equivalent to the glutamine seen in the HerA/FtsK superfamily (SM). These observations together with the statistically significant recovery of φ29-ATPases by sensitive profiles of the HerA/FtsK superfamily indicate that the former are a distinct branch of the latter superfamily. However, there were no specific features that unified φ29-like ATPases with HerA/FtsK-type packaging ATPases of other dsDNA viruses, suggesting that they are a rapidly diverging independent lineage within the HerA/FtsK superfamily.

TLSs are almost always two domain proteins with an N-terminal ASCE-type P-loop ATPase domain and a C-terminal nuclease domain with a RuvC-like version of the RNaseH fold [23, 24]. The secondary structure of the terminase ATPase domain revealed the presence of at least one additional strand after strand-2 (SM, fig. 2), placing them in a monophyletic assemblage of the ASCE division along with the HerA/FtsK, PilT, RecA and helicase superfamilies [5]. However, they lack the C-terminal β-hairpin or any other specific features characteristic of most members of the above assemblage [5] (SM). The TLS ATPase domain is distinguished from other related ATPases by the presence of a poorly conserved but universally present insert after the second β-α unit which includes strand-2. They also contain a characteristic arginine at the third position in the Walker A motif. While it could potentially act as an arginine finger in the terminase multimer, such a function remains uncertain as the arginine is absent in a few active terminases, like that of phage T1 (SM). Thus, it appears that TLS ATPase domains comprise a separate lineage within the above monophyletic assembly of the ASCE division.

The predicted adenoviral packaging ATPase (IVA2) consistently retrieved ABC ATPases as best hits. They specifically share with the ABC ATPases two polar residues at the end of the sensor-1 strand, one of which is a highly conserved histidine. Secondary structure predictions also suggest that they contain an insert with β-strands after helix-1 which might be equivalent to the corresponding insert found in all ABC ATPases [25]. Hence, the adenoviral IVA2 proteins were potentially derived from the ABC superfamily. These putative ATPases additionally contain a distinct C-terminal extension predicted to form an α+β domain with two conserved aromatic positions and several polar

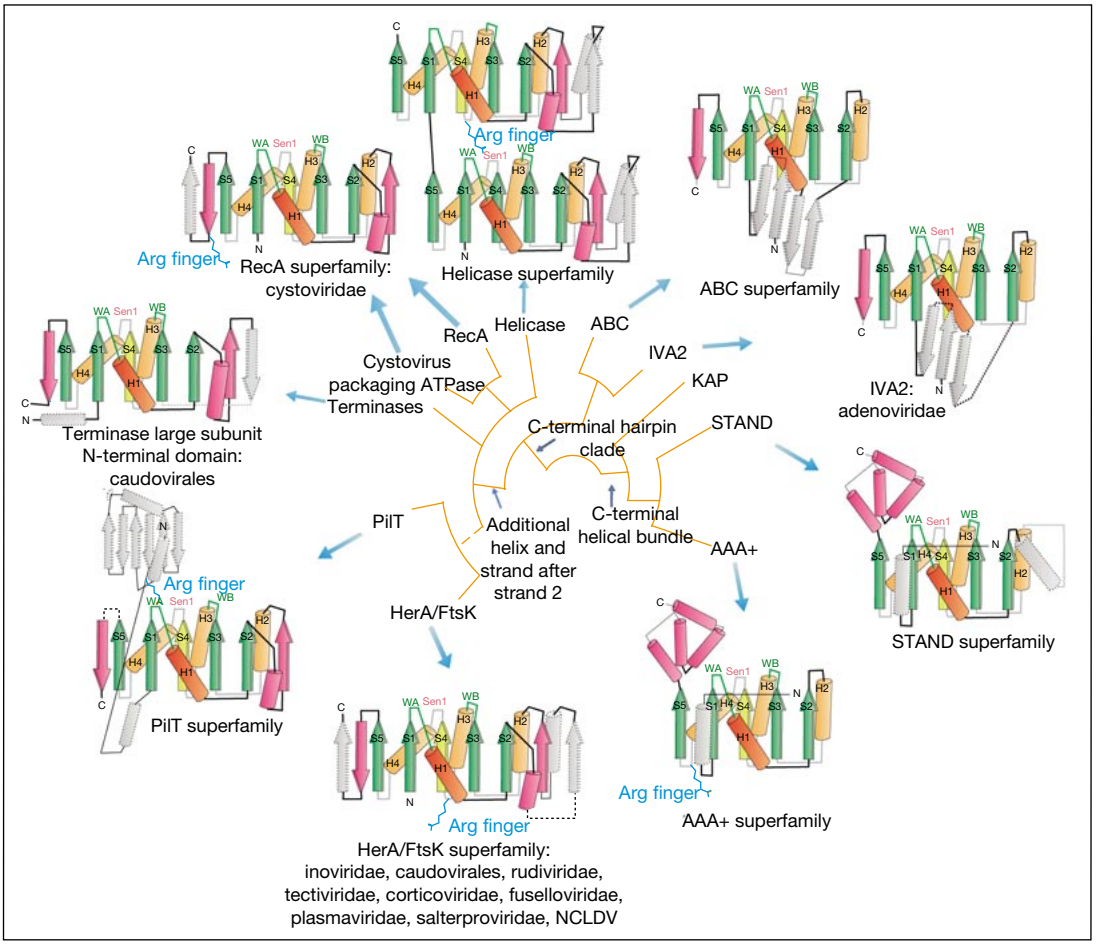


Fig. 2. Topology diagrams depicting ASCE division of P-loop NTPases and accompanying cladogram depicting higher-order relationships. Viral lineages with packaging ATPases from a specific superfamily are listed following a colon below the superfamily name. Strands and helices forming the core of the ASCE P-loop NTPase domain are numbered and colored. Strands are in green with the central strand S4 in yellow and helices in orange. Synapomorphies shared across different lineages are colored pink, elements not conserved across lineages are colored gray and outlined in broken lines. Lines connecting different lineages represent higher-order relationships constructed by comparison of shared structural and/or sequence similarities. Broken lines represent relationships with more uncertainty. Abbreviations: WA, Walker A; WB, Walker B and Sen1, sensor-1.

residues (SM). It might play a role in recognizing packaging-initiation signals in genomic regions.

Thus, it appears that DNA packaging has been derived on at least three independent occasions within the ASCE division of P-loop NTPases, with two of them being exclusively comprised of packaging ATPases (HerA/FtsK and terminase), and the third from a superfamily of ATPases that were ancestrally associated with DNA-related functions (ABC ATPases). The predicted packaging ATPases of archaeal globuloviruses can currently only be identified as members of the ASCE division with no close relationships to any of the other three classes, and might represent a fourth independent innovation. Interestingly, the only characterized packaging ATPases of dsRNA viruses, those of cystoviruses (e.g. φ12), represent another independent recruitment for packaging function from within the RecA superfamily [26, 27] (fig. 2).

Ancilliary Components of DNA Packaging Systems

Functional studies to date have not uncovered any conserved system of interacting proteins that function along with viral members of the HerA/FtsK superfamily. Previous studies have shown their cooperation with diverse nucleases in resolving target DNA during active pumping by these ATPases [5]. Hence, it is likely that these ATPases cooperate during packaging with different resolvases, including the frequently present RuvC-like resolvase, in NCLDVs and prokaryotic dsDNA viruses [19, 28, 29]. The φ-29 lineage of the HerA/FtsK superfamily appears to utilize a distinct portal protein (PP) containing a globular domain of the Src Homology 3 (SH3) fold [30]. This domain forms a multimeric ring similar to those formed by other nucleic acid binding members of this fold such as the RNA-binding Sm domain [30–32]. Adenoviruses possess a unique ancillary protein, not observed elsewhere in the viral universe, which probably functions in conjunction with the IVA2 protein to interact with genomic packaging sequences [33, 34]. Secondary structure predictions of this protein indicate a lineage-specific α-helical fold. Terminase systems show considerable diversity with different types of PPs and ancillary components like terminase small subunits (TSS). Given this diversity, we sought to investigate their origins and identify new interacting components using genomic context information.

Diversity of the Terminase-Dependent Packaging Systems: A Common Origin for Portal Proteins of All Tailed Bacteriophages

In addition to the TLS whose two domains supply ATP-dependent motor and nuclease activity, packaging systems of all characterized caudoviruses also require a PP. PPs form homo-multimers providing a conduit for DNA into the viral prohead [30, 31, 35]. In contrast to the common origin of TLSs, PPs of

these viruses were believed to belong to distinct families, typified by versions found in phage T4, T5, λ, A118 and Mu [36]. To investigate evolutionary affinities of PPs we initiated systematic transitive sequence profile searches from all known versions of PPs. As a result of these searches, we were able to recover PPs from a variety of phages or their equivalents such as the head-tail connector protein (gp8) of phage T7; consequently unifying all known PPs of tailed bacteriophages. These searches showed that every TLS-containing phage also encoded one predicted PP suggesting a strict functional association (SM). Unification of PPs of diverse phage families also implied descent from a common ancestor, just as their terminase counterparts. However, they have subsequently undergone rather drastic sequence divergence. Secondary structure prediction of the conserved core shared by PPs indicates a six-stranded region embedded between two predominantly α-helical elements. The most prominent sequence conservation is in the β-strand rich region and includes a Gxs (where 'x' is any amino acid and 's' a small residue) prior to the first conserved strand (SM). Sequence similarity-based clustering and examination of conserved shared motifs in the alignment helped us to discern eight distinct families (SM), which further grouped together into four higher order clades (T1/T5/λ-like clade, the T4/SPP1/φg1e-like clade, the phage μ-like clade and the phage T3/T7-like clade).

The presence of a conserved β-strand-rich region in PPs is reminiscent of the SH3 fold β-barrel in the φ29-type PP. The conserved Gxs motif in the former superfamily is also reminiscent of a similar motif seen in the corresponding position of SH3-like barrels [37]. Hence, despite lack of significant sequence similarity, it is not impossible that a similar β-barrel might be present in PPs of terminase-dependent systems. Likewise, herpesviral PPs, while displaying no significant sequence similarity to those of bacteriophages, also contain a core β-strand rich region suggesting the presence of a similar structure (data not shown). We propose that this β-strand rich region might form a comparable β-barrel domain, which multimerizes to give rise to the funnel shaped portal.

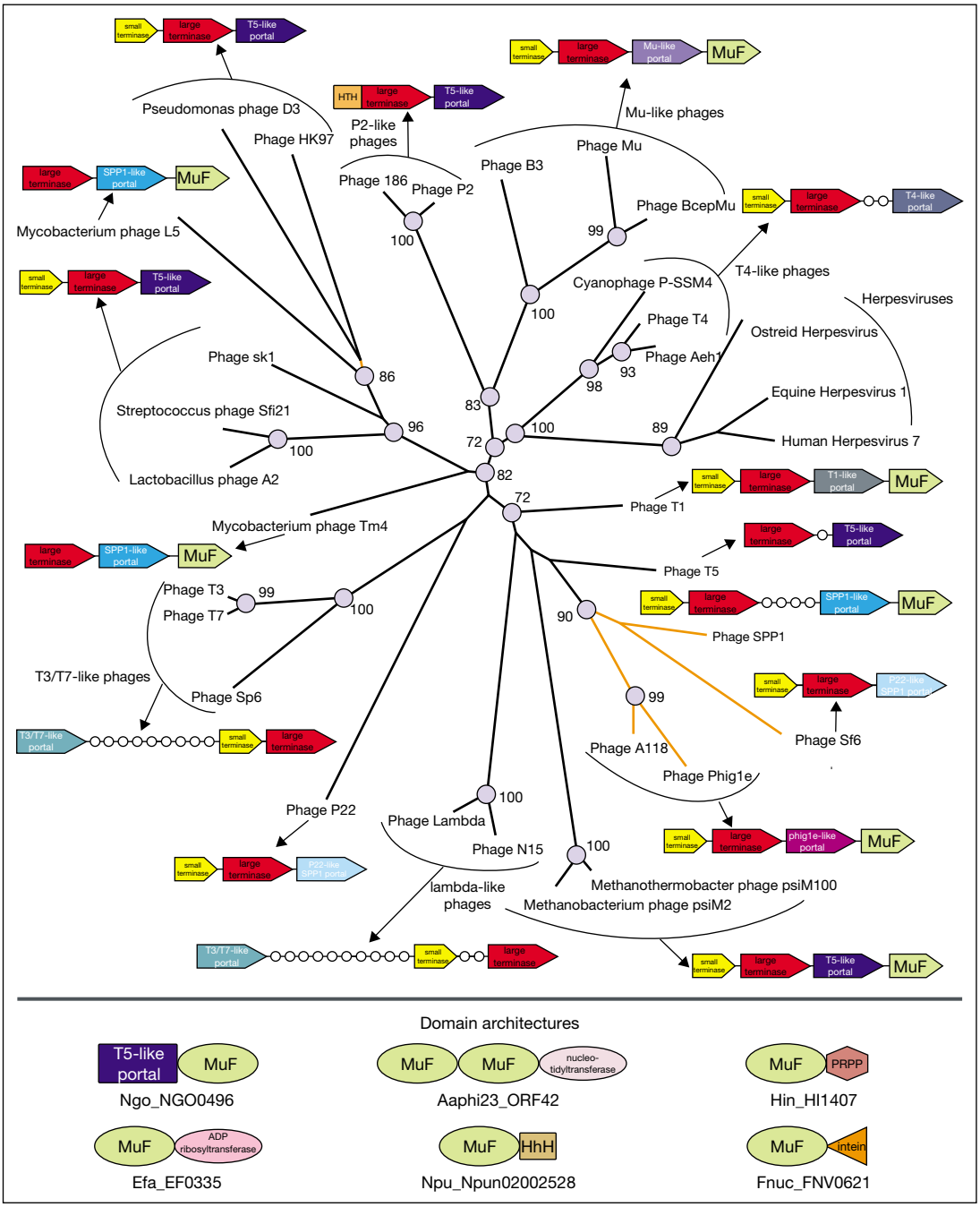
Contextual Information and Inference of Novel Components of the Terminase Portal Systems

Conserved gene neighborhoods (operons) and gene fusions have proven to be a powerful method for predicting previously unknown functional associations and protein-protein interactions in prokaryotes and their viruses [38, 39]. In order to identify other functional links to the terminase-portal system, we systematically explored all gene neighborhoods of terminase-portal pairs in bacteriophages. In terms of gene neighborhood, the most commonly found association is between the TLS and the PP, which typically occur as neighboring genes in several viral genomes (some exceptions include λ, T3/T7 and T5)

(SM, fig. 3). PP genes are rarely fused to other genes suggesting that multimerization and strict interactions with TLS are likely to select against fusion proteins. One notable fusion of the PP is with a lysozyme (e.g. *Burkholderia* prophage, gi: 78061894; fig. 3), which might correlate with the incorporation of lysozymes in viral capsids for their role in host entry.

Terminase small subunits (TSS) have been characterized in phages such as T4, T7, λ and SPP1, but corresponding small subunits have not been found in many other tailed bacteriophages [40–42]. Examination of gene neighborhoods suggested a strong association between the genes for the TSS and the TLS (fig. 3). The crystal structure of the λ small subunit shows a specialized derivative of the winged Helix-Turn-Helix (HTH) domain – the MerR-like HTH, which lacks the first of three characteristic helices of classical HTHs [43]. This suggests that the primary role of the TSS is binding DNA. Accordingly, we combined the contextual information of gene neighborhood and sequence profile searches to characterize the other TSSs and identify previously undetected versions. Our searches identified TSSs in 151 of the 206 phages containing terminase-portal systems. While all these small subunits contain the HTH fold, they included versions distinct from the MerR-type HTH seen in λ-like TSSs. In total, we identified seven distinct families of TSS and also few sporadic unclassified HTH domains. Of these, the largest families were SPP1-type TSS and D3-like TSS. The SPP1-like family was shown to contain a simple trihelical HTH module of the FIS type, while the remaining families did not belong to any previously characterized type of HTH domain and likely represent phage-specific divergent versions of the fold (SM). In a subset of phages, including P2, the SPP1-like TSS is fused to the TLS, supporting the strong functional association between the two subunits through physical interaction (SM). The above observations suggest that unlike the TLS, the TSS has been derived from the HTH fold on multiple occasions, and convergently evolved similar functional associations with the TLS.

The next major family of proteins, often encoded in the same conserved gene neighborhoods as other components of the terminase system, is the so-called MuF family. This family, typified by phage SPP1 gp7 protein, is a component of the phage prohead. In bacteriophages infecting Gram-positive bacteria, the MuF protein is known to associate with the PP and is believed to be led into the prohead by the latter [44, 45]. In our sequence profile searches, we detected MuF proteins in representatives of all major tailed prokaryotic virus families and their prophage derivatives (including one in archaeon *Methanococcus*: MJ0329). Nevertheless, several phages in each of these families lacked MuF, suggesting that it might not be an essential component of terminase-portal systems (fig. 3). The MuF gene is almost always immediately downstream of the PP gene and is associated with genes for several distinct


small terminase
large terminase
T5-like portal
Mu-like portal
MuF
HTH
Pseudomonas phage D3
Phage HK97
P2-like phages
Mu-like phages
Phage Mu
Phage B3
Phage 186
Phage P2
Phage BcepMu
T4-like portal
SPP1-like portal
Mycobacterium phage L5
100
99
Cyanophage P-SSM4
T4-like phages
Herpesviruses
Phage T4
Ostreid Herpesvirus
Phage sk1
86
Phage Aeh1
93
98
83
Equine Herpesvirus 1
Streptococcus phage Sfi21
96
100
72
89
Human Herpesvirus 7
Lactobacillus phage A2
82
T1-like portal
Phage T1
Mycobacterium phage Tm4
Phage T5
Phage T3
Phage T7
99
100
90
Phage SPP1
T3/T7-like phages
Phage Sp6
P22-like SPP1 portal
Phage Sf6
T3/T7-like portal
Phage A118
Phage Phig1e
Phage P22
Phage Lambda
100
phig1e-like portal
Phage N15
lambda-like phages
Methanothermobacter phage psiM100
Methanobacterium phage psiM2
Domain architectures
nucleo-tidyltransferase
PRPP
Ngo_NGO0496
Aaphi23_ORF42
Hin_HI1407
ADP ribosyltransferase
HhH
intein
Efa_EF0335
Npu_Npun02002528
Fnuc_FNV0621

families of portal proteins in different phages like T1, T5, Mu, λ and SPP1 (fig. 3). In one instance it is fused to a T5-like portal gene (*Neisseria* prophage, gi: 59800934), reinforcing the strong functional association between these two components.

MuF contains a characteristic C-terminal region with conserved cysteines, histidines and acidic residues suggesting it might form a distinct metal-chelating domain, which might be involved in MuF-mediated DNA binding activity. MuF proteins show a number of fusions to other domains in several (pro)phages. These include fusions to the DNA-binding Helix-hairpin-Helix (HhH) domain (*Nostoc* prophage, gi: 23130420) and several catalytic domains such as ADP ribosyltransferase (*Enterococcus* prophage, gi: 29374974), pol-β-fold nucleotidyltransferase (phage Aaφ23, gi: 31408074), PRPP amidotransferase (*Haemophilus* prophage, gi: 16273315) and multiple intein-type HINT peptidase domains (*Fusobacterium*, gi: 34763916; *Bifidobacterium*, gi: 23335596). ADP ribosyltransferases have been observed in a variety of phages, including T4 and eukaryotic NCLDVs, like PBCV and mimivirus [19]. T4 ADP ribosyltransferases ModA, ModB and Alt are packaged into phage heads, and are involved in modifying a range of host proteins [46]. Hence, the MuF might help in loading ADP-ribosyltransferase and other catalytic activities in the phage head for modification of host or viral proteins. HINT peptidases fused to MuF are related to the BUBL1 peptidase of ciliates which is involved in cleaving tandemly-fused ubiquitin repeats and ADP ribosyltransferase domains [47]. Consequently, MuF associated HINT peptidases might be similarly involved in phage head maturation. In this context, it should be noted that the portal-terminase system genes including MuF are often combined with another conserved gene neighborhood, which contains proteases involved in capsid maturation belonging to ClpP or herpesvirus assemblin-like folds [49].

Fig. 3. Phylogenetic tree of TLS depicting gene displacement among portal protein families. Phylogenetic trees were built using the least-square method with subsequent local rearrangement to obtain the maximum likelihood tree (see SM for details). Reliability of the tree topology was assessed using the RELL bootstrap method of MOLPHY, with 10,000 replications (SM). Branches where gene displacement has occurred as discussed in the text are colored orange for emphasis. Gene neighborhoods corresponding to TLSs are adjacent to branch ends, genes are shown as boxed arrows. TLS genes are colored in red, TSS colored in yellow, MuF colored in green, and PPs are colored according to family type. Nodes with bootstrap support >70% are linked by circles and labeled by bootstrap value. Domain architectures are also given below the tree, with organism abbreviations and gene names (separated by an underscore) written below. Abbreviations: HhH, helix-hairpin-helix; PRPP, PRPP amidotransferase. Please see SM for phage abbreviations.

In Situ Gene Displacement in Terminase Portal Gene Neighborhoods

Diversification of PPs into several distinct subgroups and recruitment of several distinct types of HTH domains as TSS raised the question of whether there was a correlation between distinct families of these proteins and the phylogeny of TLSs. Only TLSs show sufficient sequence conservation to reconstruct a suitably resolved phylogenetic tree through conventional methods (fig. 3). Hence, we used this tree as a reference to study the distribution of other components of the terminase-portal system and structures of their gene neighborhoods. This distribution showed the following features: (1) MuF proteins show a sporadic distribution with related phages often differing in its presence or absence. (2) Phages with related TLS might often differ in the type of PP or TSS they are associated with. For example, phage SPP1 has an SPP1-like PP (PP2 family) while the related phage Sf6 contains a P22-like version. Likewise, related TLSs of phages P2 and B3 differ in terms of associated TSS and PP and presence or absence of MuF (fig. 3).

These observations suggest that terminase-portal gene neighborhoods are prone to: (1) frequent gene loss and acquisition, evidenced by sporadic distribution of MuF and (2) in situ displacement of functionally equivalent proteins by evolutionarily unrelated or distantly related counterparts. This situation is parallel to previously observed gene neighborhoods of phage single strand annealing proteins and capsid maturation proteases [48, 49]. Presence of relatively strict gene orders (TSS followed by TLS, PP and MuF) suggests strong constraints with respect to their synthesis and interactions. General rarity or absence of gene fusions among TSS, TLS and PP suggest that their interactions are strongly coupled without much scope for additional associations. Based on gene order and nature of domain fusions, we speculate that TSS is synthesized first and associates with viral DNA. It subsequently recruits the TLS which processes DNA and recruits the PP through which DNA is loaded into the prohead. The PP in turn appears to recruit MuF, which might help position DNA into proheads and recruit other catalytic activities for capsid maturation.

Evolutionary Considerations and General Conclusions

The systematic survey of diverse active viral DNA-packaging systems suggests that their motors have been derived from two major superfamilies of ASCE ATPases (HerA/FtsK and TLS N-terminal domain). The remaining packaging motors are also derived from the ASCE division, but are very limited in their spread and appear to lack an extended evolutionary history. Taken together with the monophyly of capsid proteins of several DNA and RNA viruses, this suggests an early origin for the two major ATP-dependent DNA-packaging systems in the context of ancient pre-existing capsid-like envelopes [19].

Interestingly, both ancient superfamilies of packaging ATPases function in conjunction with DNAses that process or manipulate the products of genome replication. While TLSs contain the C-terminal RNaseH fold nuclease domain, the HerA/FtsK superfamily functions with a range of distinct nucleases in cellular and viral systems, such as XerC/XerD, NurA, RCR, pT181/Rep, Sir2 and possibly RuvC-like resolvases (in several NCLDVs) [5, 28, 29]. The RNaseH-fold domain in TLS is most closely related in terms of its conserved active site to RuvC resolvases and nuclease domains of several transposases (such as TnpA, Mariner, Hermes, Rag1/Transib and retroviral integrases) [24, 50, 51]. Thus, ATPases of both packaging systems probably associated with an ancestral DNA manipulating nuclease of the RNaseH fold, which appears to have diversified into nuclease, integrase or resolvase families of viral and cellular replicons. HerA/FtsK ATPases form ring-structures and lack domain fusions with their nuclease partners. This appears to have allowed more frequent evolutionary displacements of their nuclease partners by functionally equivalent nucleases [5]. In contrast, there is no evidence for TLSs forming comparable arginine finger-stabilized rings, and fusion with their nuclease partner appears to have been retained throughout their evolution. In general, functional associations between nucleases and packaging ATPases suggest that from inception packaging systems were closely associated with post-replication genome segregation. Increasing size of DNA-based genomes probably provided the selection pressure for emergence of such systems [19].

Interestingly, diversification of several other superfamilies in the ASCE division of P-loop NTPases might be linked to emergence and expansion of DNA-based replication systems. These include DNA helicases of AAA+, recombinases of RecA, and higher order chromosome condensation proteins of ABC superfamilies. Hence, the two major DNA packaging systems probably arose as part of this diversification of ASCE NTPases concomitant with diversification of DNA-based replicons that occurred well before the emergence of the Last Universal Common Ancestor (LUCA) of cellular life [5]. The nature of the envelope of early replicons, lipid membranes or purely protein capsids, appears to have played a principal role in emergence of the two independent packaging motors. In this context, it is notable that cellular systems (bacteria and archaea) use packaging ATPases related to those of viruses with lipid inner membranes. Thus, precursors of cellular compartments could have emerged from systems similar to lipid containing viral capsids [19]. Thus, it appears that the precursors of the principal packaging ATPases of viral systems and cellular chromosome-pumping ATPases, like FtsK appear to have emerged during the pre-LUCA radiation of the ASCE clade, and followed independent history ever since. While both major packaging systems remained largely mutually exclusive in viruses, on rare occasions we do find potential hybrid systems. The φ29 system uses a

HerA/FtsK ATPase but depends on a PP analogous to caudoviruses. Like the latter, it lacks an inner membrane, and has a unique hexameric RNA component (prohead RNA or pRNA) [52]. It remains unclear if this pRNA is a remnant of a more ancient system or merely a lineage-specific innovation of φ29-like phages. Similarly, evolution of adenoviruses might have involved displacement of the HerA/FtsK ATPase of the above-mentioned Tlr-like DNA transposons by a neomorphic packaging system. The adenovirus IVA protein, as well as RecA-like packaging ATPases limited to cystoviruses, could have emerged via rapid divergence from either viral or cellular precursors. Our unification of PPs suggests that the terminase-dependent system deployed PPs from the earliest stages of their existence. The observation that most of these viruses also contain a version of the HTH domain (TSS) suggests that there might have been a third component that recruited the motor to DNA even in ancestral versions of this system.

We hope this overview might help in further experimental investigations on functional interactions in these systems.

Acknowledgements

The authors gratefully acknowledge the Intramural Research Program of the National Institutes of Health, USA for funding their research.

Supplementary Material

The supplementary material can be accessed from ftp://ftp.ncbi.nih.gov/pub/aravind/portal/.

Note Added in Proof

While this manuscript was being processed for publication, the crystal structure of the TLS was solved [53] and was shown to belong to the ASCE division as predicted. The presence of a helix strand unit after strand-2 reaffirmed its position with respect to other ATPases (fig. 2). The C-terminal region, however, is much diverged from other ATPases. The crystal structure also confirms the presence of an arginine finger in Walker A as predicted.

References

1 Wagner KE, Hewlett MJ: Basic Virology, ed 2. Blackwell Publishers, Oxford, 2003.
2 Catalano CE: Viral Genome Packaging: Genetics, Structure, and Mechanism, Kluwer Academic/Plenum publisher, New York, 2005.

3 Rao AL: Genome packaging by spherical plant RNA viruses. Annu Rev Phytopathol 2006;44: 61–87.
4 Bernal RA, Hafenstein S, Esmeralda R, Fane BA, Rossmann MG: The phiX174 protein J mediates DNA packaging and viral attachment to host cells. J Mol Biol 2004;337:1109–1122.
5 Iyer LM, Makarova KS, Koonin EV, Aravind L: Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. Nucleic Acids Res 2004;32:5260–5279.
6 Stromsten NJ, Bamford DH, Bamford JKH: In vitro DNA packaging of PRD1: a common mechanism for internal-membrane viruses. J Mol Biol 2005;348:617–629.
7 Nandhagopal N, Simpson AA, Gurnon JR, Yan X, Baker TS, et al: The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus. Proc Natl Acad Sci USA 2002;99: 14758–14763.
8 Hendrix RW: Evolution: the long evolutionary reach of viruses. Curr Biol 1999;9:914–917.
9 Newcomb WW, Juhas RM, Thomsen DR, Homa FL, Burch AD, et al: The UL6 gene product forms the portal for entry of DNA into the herpes simplex virus capsid. J Virol 2001;75:10923–10932.
10 Catalano CE: The terminase enzyme from bacteriophage lambda: a DNA-packaging machine. Cell Mol Life Sci 2000;57:128–148.
11 Black LW: DNA packaging and cutting by phage terminases: control in phage T4 by a synaptic mechanism. Bioessays 1995;17:1025–1030.
12 Rentas FJ, Rao VB: Defining the bacteriophage T4 DNA packaging machine: evidence for a C-terminal DNA cleavage domain in the large terminase/packaging protein gp17. J Mol Biol 2003;334:37–52.
13 Goetzinger KR, Rao VB: Defining the ATPase center of bacteriophage T4 DNA packaging machine: requirement for a catalytic glutamate residue in the large terminase protein gp17. J Mol Biol 2003;331:139–154.
14 Ibarra B, Valpuesta JM, Carrascosa JL: Purification and functional characterization of p16, the ATPase of the bacteriophage Phi29 packaging machinery. Nucleic Acids Res 2001;29:4264–4273.
15 Koonin EV, Senkevich TG, Chernos VI: Gene A32 product of vaccinia virus may be an ATPase involved in viral DNA packaging as indicated by sequence comparisons with other putative viral ATPases. Virus Genes 1993;7:89–94.
16 Iyer LM, Aravind L, Koonin EV: Common origin of four diverse families of large eukaryotic DNA viruses. J Virol 2001;75:11720–11734.
17 Wuitschick JD, Gershan JA, Lochowicz AJ, Li S, Karrer KM: A novel family of mobile genetic elements is limited to the germline genome in *Tetrahymena thermophila*. Nucleic Acids Res 2002;30:2524–2537.
18 Zhang W, Imperiale MJ: Requirement of the adenovirus IVa2 protein for virus assembly. J Virol 2003;77:3586–3594.
19 Iyer LM, Balaji S, Koonin EV, Aravind L: Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res 2006;117:156–184.
20 Leipe DD, Wolf YI, Koonin EV, Aravind L: Classification and evolution of P-loop GTPases and related ATPases. J Mol Biol 2002;317:41–72.
21 Iyer LM, Leipe DD, Koonin EV, Aravind L: Evolutionary history and higher order classification of AAA+ ATPases. J Struct Biol 2004;146:11–31.
22 Neuwald AF, Aravind L, Spouge JL, Koonin EV: AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. Genome Res 1999;9: 27–43.
23 Kanamaru S, Kondabagil K, Rossmann MG, Rao VB: The functional domains of bacteriophage t4 terminase. J Biol Chem 2004;279:40795–40801.
24 Ponchon L, Boulanger P, Labesse G, Letellier L: The endonuclease domain of bacteriophage terminases belongs to the resolvase/integrase/ribonuclease H superfamily: a bioinformatics analysis validated by a functional study on bacteriophage T5. J Biol Chem 2006;281:5829–5836.
25 Holland IB, Blight MA: ABC-ATPases, adaptable energy generators fuelling transmembrane movement of a variety of molecules in organisms from bacteria to humans. J Mol Biol 1999;293:381–399.
26 Lisal J, Kainov DE, Bamford DH, Thomas GJ Jr, Tuma R: Enzymatic mechanism of RNA translocation in double-stranded RNA bacteriophages. J Biol Chem 2004;279:1343–1350.

27 Kainov DE, Pirttimaa M, Tuma R, Butcher SJ, Thomas GJ Jr, et al: RNA packaging device of double-stranded RNA bacteriophages, possibly as simple as hexamer of P4 protein. J Biol Chem 2003;278:48084–48091.
28 Garcia AD, Aravind L, Koonin EV, Moss B: Bacterial-type DNA holliday junction resolvases in eukaryotic viruses. Proc Natl Acad Sci USA 2000;97:8926–8931.
29 Aravind L, Makarova KS, Koonin EV: SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. Nucleic Acids Res 2000;28:3417–3432.
30 Simpson AA, Tao Y, Leiman PG, Badasso MO, He Y, et al: Structure of the bacteriophage phi29 DNA packaging motor. Nature 2000;408:745–750.
31 Guasch A, Pous J, Parraga A, Valpuesta JM, Carrascosa JL, Coll M: Crystallographic analysis reveals the 12-fold symmetry of the bacteriophage phi29 connector particle. J Mol Biol 1998;281: 219–225.
32 Mura C, Phillips M, Kozhukhovsky A, Eisenberg D: Structure and assembly of an augmented Sm-like archaeal protein 14-mer. Proc Natl Acad Sci USA 2003;100:4539–4544.
33 Perez-Romero P, Gustin KE, Imperiale MJ: Dependence of the encapsidation function of the adenovirus L1 52/55-kilodalton protein on its ability to bind the packaging sequence. J Virol 2006;80: 1965–1971.
34 Gustin KE, Lutz P, Imperiale MJ: Interaction of the adenovirus L1 52/55-kilodalton protein with the IVa2 gene product during infection. J Virol 1996;70:6463–6467.
35 Bazinet C, Benbasat J, King J, Carazo JM, Carrascosa JL: Purification and organization of the gene 1 portal protein required for phage P22 DNA packaging. Biochemistry 1988;27:1849–1856.
36 Mitchell MS, Matsuzaki S, Imai S, Rao VB: Sequence analysis of bacteriophage T4 DNA packaging/terminase genes 16 and 17 reveals a common ATPase center in the large subunit of viral terminases. Nucleic Acids Res 2002;30:4009–4021.
37 Anantharaman V, Aravind L: Novel conserved domains in proteins with predicted roles in eukaryotic cell-cycle regulation, decapping and RNA stability. BMC Genomics 2004;5:45.
38 Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res 2001;11: 356–372.
39 Huynen M, Snel B, Lathe W 3rd, Bork P: Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 2000;10:1204–1210.
40 Gual A, Alonso JC: Characterization of the small subunit of the terminase enzyme of the *Bacillus subtilis* bacteriophage SPP1. Virology 1998;242:279–287.
41 Lin H, Simon MN, Black LW: Purification and characterization of the small subunit of phage T4 terminase, gp16, required for DNA packaging. J Biol Chem 1997;272:3495–3501.
42 Bain DL, Berton N, Ortega M, Baran J, Yang Q, Catalano CE: Biophysical characterization of the DNA binding domain of gpNu1, a viral DNA packaging protein. J Biol Chem 2001;276: 20175–20181.
43 de Beer T, Fang J, Ortega M, Yang Q, Maes L, et al: Insights into specific DNA recognition during the assembly of a viral genome packaging machine. Mol Cell 2002;9:981–991.
44 Stiege AC, Isidro A, Droge A, Tavares P: Specific targeting of a DNA-binding protein to the SPP1 procapsid by interaction with the portal oligomer. Mol Microbiol 2003;49:1201–1212.
45 Droge A, Santos MA, Stiege AC, Alonso JC, Lurz R, et al: Shape and DNA packaging activity of bacteriophage SPP1 procapsid: protein components and interactions during assembly. J Mol Biol 2000;296:117–132.
46 Depping R, Lohaus C, Meyer HE, Ruger W: The mono-ADP-ribosyltransferases Alt and ModB of bacteriophage T4: target proteins identified. Biochem Biophys Res Commun 2005;335: 1217–1223.
47 Dassa B, Yanai I, Pietrokovski S: New type of polyubiquitin-like genes with intein-like autoprocessing domains. Trends Genet 2004;20:538–542.
48 Iyer LM, Koonin EV, Aravind L: Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52. BMC Genomics 2002;3:8.
49 Liu J, Mushegian A: Displacements of prohead protease genes in the late operons of double-stranded-DNA bacteriophages. J Bacteriol 2004;186:4369–4375.

50 Rice PA, Baker TA: Comparative architecture of transposase and integrase complexes. Nat Struct Biol 2001;8:302–307.
51 Kapitonov VV, Jurka J: RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 2005;3:e181.
52 Xiao F, Moll WD, Guo S, Guo P: Binding of pRNA to the N-terminal 14 amino acids of connector protein of bacteriophage phi29. Nucleic Acids Res 2005;33:2640–2649.
53 Sun S, Kondabagil K, Gentz PM, Rossmann MG, Rao VB: The structure of the ATPase that powers DNA packaging into bacteriophage T4 procapsids. Mol Cell 2007;25:943–949.

L. Aravind
National Center for Biotechnology Information,
National Library of Medicine,
National Institutes of Health,
Bethesda, MD 20894, USA
Tel (301) 594-2445, Fax (301) 480-9241, E-Mail aravind@ncbi.nlm.nih.gov

Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 66–80

General Trends in the Evolution of Prokaryotic Transcriptional Regulatory Networks

M. Madan Babu[a,b], *S. Balaji*[a], *L. Aravind*[a]

[a]National Center for Biotechnology Information, National Institutes of Health, Bethesda, Md., USA; [b]MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK

Abstract

Gene expression in organisms is controlled by regulatory proteins termed transcription factors, which recognize and bind to specific nucleotide sequences. Over the years, considerable information has accumulated on the regulatory interactions between transcription factors and their target genes in various model prokaryotes, such as *Escherichia coli* and *Bacillus subtilis*. This has allowed the representation of this information in the form of a directed graph, which is commonly referred to as the transcriptional regulatory network. The network representation provides us with an excellent conceptual framework to understand the structure of the transcriptional regulation, both at local and global levels of organization. Several studies suggest that the transcriptional network inferred from model organisms may be approximated by a scale-free topology, which in turn implies the presence of a relatively small group of highly connected regulators (hubs or global regulators). While the graph theoretical principles have been applied to infer various properties of such networks, there have been few studies that have actually investigated the evolution of the transcriptional regulatory networks across diverse organisms. Using recently developed computational methods that exploit various evolutionary principles, we have attempted to reconstruct and compare these networks across a wide-range of prokaryotes. This has provided several insights on the modification and diversification of network structures of various organisms in course of evolution. Firstly, we observed that target genes show a much higher level of conservation than their transcriptional regulators. This in turn suggested that the same set of functions could be differently controlled across diverse organisms, contributing significantly to their adaptive radiations. In particular, at the local level of network structure, organism-specific optimization of the transcription network has evolved primarily via tinkering of individual regulatory interactions rather than whole scale reuse or deletion of network motifs (local structure). In turn, as phylogenetic diversification proceeds, this process appears to have favored repeated convergence to scale-free-like structures, albeit with different regulatory hubs.

Copyright © 2007 S. Karger AG, Basel

The pioneering studies by Jacob and Monod suggested the existence of regulatory proteins which bind to DNA elements upstream of other genes and control their expression. These regulatory proteins, termed transcription factors, respond to different signals and in turn activate or repress the expression of their target genes at appropriate instances [1–5]. Following these findings, several studies over the many years have accumulated a wealth of information on individual regulatory interactions mediated by these transcription factors in various model organisms [6–13]. More recently, there has been considerable interest and effort to assemble this information to derive what is termed the transcriptional regulatory network of an organism [14–16]. The topology of the transcriptional regulatory network is best modeled as a network or a graph with nodes representing transcription factors and target genes, and directed edges connecting the former to the latter [15–17].

Several recent studies on transcriptional networks of prokaryotes and eukaryotes have shown that the structure of such networks shows three levels of organization [15]. At the most basic level, the network contains individual regulatory interactions between transcription factors and targets (fig. 1). At the intermediate level, studies have shown multiple basic units to be organized into functionally distinct units called network motifs (fig. 1). Different types of motifs are defined based on patterns of interconnections between the basic units, and multiple copies of individual motif-types are found in different contexts within the network. Finally, at the highest level of organization, the set of all transcriptional regulatory interactions in a cell form the global structure and has been shown to have a hierarchical or a scale-free topology. In other words, such a global structure is characterized by the presence of a majority of transcription factors which regulate few genes and the presence of a few transcription factors, called the regulatory hubs, which regulate many genes (fig. 1).

While there has been significant progress, due to several experimental studies performed over many years, in unraveling the transcriptional regulatory networks of various model organisms such as *E. coli* and *B. subtilis*, much less information is available on the transcriptional networks of other prokaryotes. In order to gain a better understanding of the transcriptional regulatory network in other organisms, computational methods to extrapolate this information from model organisms to poorly studied organisms by exploiting the wealth of information in the form of publicly available completely sequenced genomes have been developed [18–29]. Such methods can be broadly grouped into two classes: (i) Orthology based methods: This approach exploits the basic principle that orthologous transcription factors regulate orthologous target genes in the different genomes. This method requires the transcriptional regulatory network for a reference organism and uses the protein sequences of transcription factors and target genes in the reference network to identify orthologs in the query

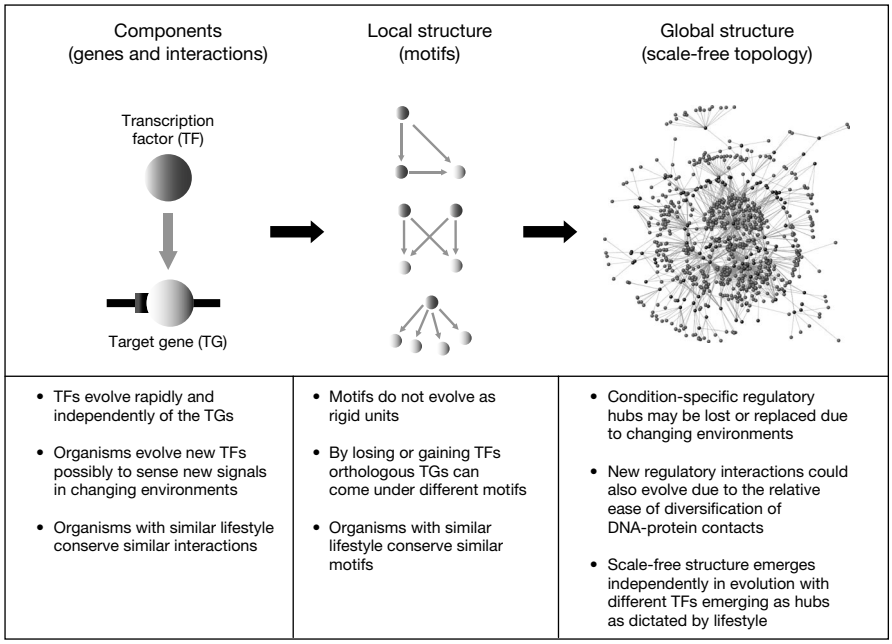

Components
(genes and interactions)
Local structure
(motifs)
Global structure
(scale-free topology)
Transcription
factor (TF)
Target gene (TG)
• TFs evolve rapidly and independently of the TGs
• Organisms evolve new TFs possibly to sense new signals in changing environments
• Organisms with similar lifestyle conserve similar interactions
• Motifs do not evolve as rigid units
• By losing or gaining TFs orthologous TGs can come under different motifs
• Organisms with similar lifestyle conserve similar motifs
• Condition-specific regulatory hubs may be lost or replaced due to changing environments
• New regulatory interactions could also evolve due to the relative ease of diversification of DNA-protein contacts
• Scale-free structure emerges independently in evolution with different TFs emerging as hubs as dictated by lifestyle

Fig. 1. Structure and evolution of transcriptional regulatory networks in prokaryotes. There are three levels of organization of network structure. (i) The basic unit is made up of a transcription factor, its target gene and a regulatory interaction represented as a directed arrow. (ii) At the local level, the basic unit forms network motifs, which are small patterns of interconnections with specific information processing ability. (iii) The set of all regulatory interactions in an organism, which is a representation of the transcriptional program of a cell, is referred to as the global structure of the network. The observed general evolutionary trends at the three levels of organizations of the network suggest that transcriptional regulatory networks in prokaryotes are very flexible and rapidly adapt to changes in environment by tinkering individual interactions to arrive at an organism specific optimal design.

organism in order to infer regulatory interactions in the genome of interest. (ii) Binding site profile based methods: This approach requires reliable information on the DNA binding site for a transcription factor. Further, it exploits the fact that presence of the same binding site upstream of different genes in a closely related species would imply a regulatory influence of an orthologous transcription factor on the expression of the nearby gene through the same binding site. Both these methods have their advantages and disadvantages; the former method allows prediction of conserved interactions and loss of interactions in distantly related organisms but does not facilitate discovery of novel targets for a given transcription factor. In contrast, the latter method allows detection of

novel targets for a transcription factor but is not applicable to distantly related genomes because DNA regulatory elements are shorter and evolve much faster than the protein-coding sequences, hence making detection of new interactions unreliable.

The availability of complete genome sequences of over 300 prokaryotes and the understanding of the structure of transcriptional regulatory networks have allowed us to address several fundamental questions on the origins and evolution of the transcriptional regulatory networks. In addition, the availability of the previously discussed methods has provided us with an opportunity to identify the distinct evolutionary trends in shaping of transcriptional regulatory networks at various levels of organization. In this chapter, we review the results from recent studies which have addressed these questions at various levels of resolution and present a summary of the general trends that can be discerned.

Evolution of Transcription Factors and Target Genes

Regulatory interactions between transcription factors and target genes could potentially evolve through two distinct modes: (i) in which both the transcription factor and the target gene co-evolve, i.e. present or absent as a pair or (ii) the transcription factor and the target gene evolve independently of each other. Our analysis of the conservation patterns (employing the orthology based method) of the genes and the regulatory interactions across 175 different prokaryotes revealed several interesting trends [29].

Using the *E. coli* transcriptional regulatory network, which consisted of 112 transcription factors, 755 target genes and 1,292 regulatory interactions, as the reference network, we found that the evolutionary retention of transcription factors in other organisms is lower than their target genes. The relatively low retention of transcription factors in other organisms does not mean that the regulatory influences on the more highly retained target genes are absent. We found that each organism has evolved its own set of transcriptional regulators that are not orthologous to other proteins suggesting innovation of regulatory proteins, possibly to sense new signals in changing environments and provide a new set of regulatory influences on target genes. Thus evolutionary forces appear to independently retain or discard transcription factors and their targets, with a higher frequency of loss or replacement of the former.

Several studies along these lines have recently demonstrated that there is a non-linear increase in the number of transcription factors encoded as the genome size of the organism increases [30–32]. These observations suggest that as genome size increases, more transcription factors are needed to regulate specialized groups of genes individually. Alternatively, it may also suggest the need

to integrate distinct inputs in order to introduce more layers in the regulatory hierarchy of metabolically or organizationally complex organisms with large genomes. These studies also revealed that (i) in parasites with small genomes, transcription factors have been lost due to absence of selective pressure for regulating target genes. These parasites could depend on the host cellular machinery for the fulfillment of roles performed by some of their proteins and (ii) in larger genomes, target genes are often controlled by additional regulators or regulators that are non-orthologous, so that there is an integration of a variety of different inputs that are typically dependent on the environmental niche of the organism [29, 32, 33]. For instance, in a number of phylogenetically distant free-living bacteria, including proteobacteria, *Bacillus subtilis* and *Streptomyces*, there is an expansion of the so-called one-component transcription factors of the LysR family, which sense a wide range of small molecule ligands. However, there are no homologs of such transcription factors in any of their close relatives, which are obligate pathogens. This is consistent with the need to sense a similar set of environmental metabolites by all of the above-mentioned free-living bacteria.

At the level of regulatory interactions, we found that organisms which are phylogenetically distantly related but share similar environmental life-style tend to significantly conserve regulatory interactions, hinting a prominent influence for environment or life-style in selecting for their conservation. For example, bacteria with comparable genome size, such as several species of *Bacillus*, *Corynebacterium* and *Mycobacterium*, whose principal habitat is the soil, conserve orthologous regulatory interactions. Likewise, the obligate or intracellular parasites from diverse bacterial clades, namely *Mycoplasma*, *rickettsiae* and *chlamydiae*, conserve similar regulatory interactions. To test the generality of this observation, an index which measures similarity in network structure and lifestyle (LSI) was developed [29] which revealed the existence of a strong evolutionary trend: organisms belonging to the same lifestyle have a significantly higher number of regulatory interactions in common in comparison to organisms from other lifestyle classes.

In addition, analysis of the conserved regulatory interactions in the different genomes allowed us to speculate about the components of the ancestral networks in the different phylogenetic lineages. The common ancestor of archaea and eubacteria appears to have had quite a few global regulatory proteins (with more than 14 target genes) such as Crp (cAMP receptor protein/regulator), Fnr (fumarate nitrate reduction regulator) and Lrp (regulator for leucine regulon). The predicted ancestral network at this level might have contained up to 62 TFs that regulate genes required for basic processes that sustain life, which include regulators for genes involved in purine biosynthesis, fructose utilization, xylulose utilization, iron uptake, fatty acid biosynthesis, anaerobic respiration and

amino acid biosynthesis. There has been gain of specific regulatory systems in course of eubacterial evolution. These include transcriptional regulators that can sense a variety of different sugar molecules and their target genes that utilize these sugars (e.g. mellibiose, mannitol, glucitol, galactose, etc.) to generate energy. Within particular bacterial lineages, e.g. the firmicute lineage, which includes the endospore-forming bacteria and actinobacteria, appears to have lost regulatory systems that can utilize L-idonate, and the sulphur utilization system. Further, in the proteobacteria and cyanobacteria, there have been multiple instances where various regulatory systems have been lost in the different lineages. These results point to the possibility that several regulators were present in the ancestral genome but have been lost, displaced or retained as organisms colonized and adapted to new environmental niches.

Evolution of the Local Network Structure

Regulatory networks can be fragmented into fundamental regulatory subsystems or motifs which when put together reconstruct the entire network. In the case of the *E. coli* network, three types of motifs (1) feed-forward motif, (2) single input motif and (3) multiple-input motifs have been discerned through a combination of computational and a series of experimental studies aimed at understanding their functions [34–39]. Such studies have elucidated that the feed-forward motif could ensure regulation of target genes only when a persistent signal is received, thereby filtering noise or fluctuation in the input signal. It was also demonstrated that the single input motif could co-ordinate global changes in gene expression and could enforce an order in the patterns of gene expression of its targets and that the multiple input motif could integrate different signals and hence could differentially regulate the target genes.

In principle evolution of these network motifs could follow any of the following trajectories: (i) a trend where all interactions in a network motif are conserved in other organisms or (ii) a trend where motifs are not conserved as complete units due to which individual interactions may be lost or gained during the course of evolution. Given that the network motifs have specific information processing ability, one might expect these motifs to be conserved as relatively rigid units (all the components of the motifs are conserved) once they have emerged. However, analysis of the conservation patterns of these network motifs across the 175 genomes [29] revealed the contrarily that network motifs are not conserved as complete units in other organisms. At a first glance, it was surprising to find that organisms which were evolutionary close did not conserve regulatory network motifs whereas several organisms which were distantly related conserved orthologous network motifs (fig. 2a). For example, Fnr

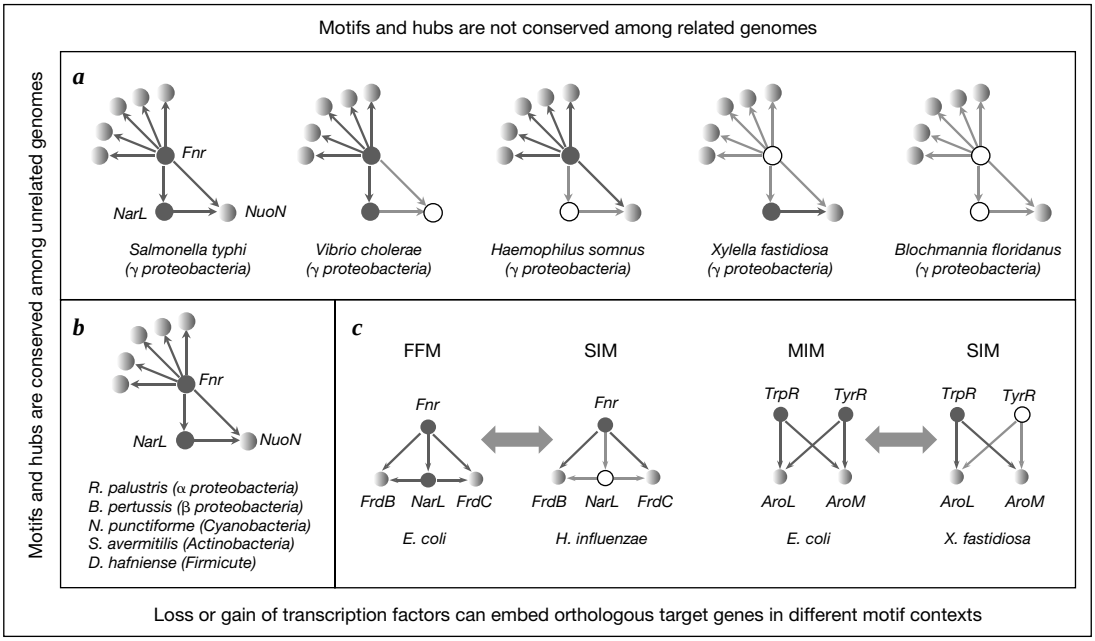


Fig. 2. Evolution of network motifs. ***a*** A feed-forward motif formed by transcription factors Fnr, NarL and NuoN in *E. coli* is completely conserved in a closely related genome, *Salmonella typhi*, but not in other gamma-proteobacterial genomes. ***b*** Distantly related organisms that have preserved all interactions in the regulatory motif and that have conserved the regulatory hub, Fnr. ***c*** Analysis of partially conserved motifs revealed that by losing (or gaining) specific transcription factors, orthologous genes in different genomes could be embedded in different motif contexts. Thus evolution tinkers with specific regulatory interactions when orthologous genes in organisms living in a different environment need to be expressed differently. In these figures the TFs and TGs are represented by dark grey circles and light grey circles, respectively, while white circles denote their absence.

(a global regulator, activated during low oxygen levels), NarL (transcriptional regulator of a two-component signal transduction system) and NuoN (subunit of the NADH dehydrogenase complex I) form a feed-forward motif in *E. coli*, which is not completely conserved as a unit in other gamma proteobacterial genomes. In contrast, all the interactions in this motif are conserved in several distantly related genomes such as the beta-proteobacterium *B. pertussis* and the firmicute *D. hafniense* (fig. 2b). Further careful analysis revealed the role of the environment in shaping the structure of these network motifs. It was found that in instances where the network motifs were not conserved in closely related organisms, they had significant differences in their life-styles. Strikingly, in

instances where distantly related organisms were found to conserve regulatory network motifs, it was seen that they had a considerable similarity in their lifestyle. A comprehensive analysis to assess the generality of this observation where organisms were grouped according to their lifestyle similarity and assessed for similarity in their network motif content revealed a statistically significant trend that organisms with similar lifestyle tend to regulate their target genes by means of similar network motifs [27].

It was not immediately clear how local network structure or network motifs of organisms evolve with diversification of life-style or environments. A case by case analysis of some of the partially conserved regulatory network motifs of organisms living in different environments revealed that by losing or gaining individual transcription factors, orthologous target genes could be potentially expressed in different ways according to the requirements of the organism. This meant that by losing or gaining individual regulatory proteins, organisms living in different environments can regulate orthologous target genes through different network motifs [29]. For instance, genes which are regulated through a feed-forward motif (FFM) can be regulated as a part of a single input motif (SIM) by losing a transcription factor (fig. 2c). Note that the regulation of a gene through a FFM would ensure that the target gene expression is not sensitive to fluctuations in input signals. Whereas regulation through a SIM would ensure expression of target genes as long as there is some input signal, for instance, the presence of a particular metabolite. Likewise, target genes regulated through a multiple input motif (MIM) in one organism can come to be regulated through a single input motif (SIM) by losing one of the transcription factors (fig. 2c).

For example, in *E. coli*, which is adapted to a lifestyle with largely fixed aerobic and anaerobic phases, the fumarate reductase genes (FrdB and FrdC, which converts fumarate to succinate under anaerobic conditions to derive energy) are not expressed unless there is a persistent signal for lack of oxygen received through a feed forward motif involving both Fnr and NarL. In contrast, *Haemophilus influenzae*, which encounters rapid redox fluctuations during host infection and needs to regulate the fumarate reductase genes more quickly than *E. coli*, appears to depend solely on Fnr for the response by employing a single input motif. Thus by losing a transcription factor, genes that are tightly regulated through a FFM can be regulated in a much simpler manner (fig. 2c).

These observations reveal an important principle in the evolution of network motifs that orthologous genes in related organisms living in different environments may acquire distinct patterns of gene expression by embedding them in appropriate motif context in order to adapt better to changing environments [40–44, 29]. Our findings also indicate that different organisms arrive at the best possible solutions to regulate the same gene by tinkering specific regulatory interactions in order to optimize expression levels rather than by duplicating

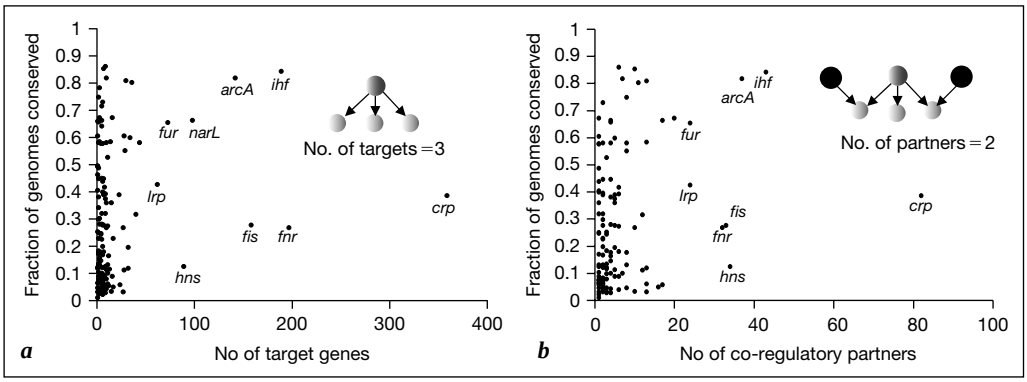


Fig. 3. Conservation v/s connectivity plot for the transcription factors in the transcriptional network of *E. coli*. ***a*** Fraction of the 330 genomes conserved against the number of target genes for a transcription factor. The connectivity in terms of the number of target genes for the transcription factor shown in black is three. ***b*** Fraction of the genomes conserved against the number of regulatory partners for a given transcription factor. The number of co-regulatory proteins for the transcription factor shown in black is two as it co-regulates its target genes with two other transcription factors. arcA is anaerobic respiration regulatory protein, ihf is integral host factor and fur is ferric uptake regulator.

groups of genes which are already a part of a motif. In this context, other studies on duplicated genes within the transcriptional regulatory network of *E. coli* and yeast [45–47] have shown that network motifs have not evolved by duplication of complete ancestral motifs lending support to our interpretation that the same interactions, which is a part of a motif in one organism, could have existed in different regulatory contexts in their ancestral genomes.

Evolution of the Global Network Structure

Results from graph theoretical studies and the fact that global regulatory hubs control the expression of several genes suggested that such hubs would assume importance in transcriptional networks and hence be more conserved in evolution than other transcription factors. In our investigation on the evolution of the global structure of prokaryotic transcriptional regulatory networks [29], we observed that transcriptional regulatory hubs are not preferentially retained than any other transcription factors in the network (fig. 3a). Careful analysis of the transcription factors which were lost in other organisms revealed that these are regulatory hubs that were largely condition specific (narL, crp, etc.) and

hence were lost in instances where the organism would not experience such a condition. For instance, an organism which has been adapted to live in an aerobic environment could dispense global regulators which are required only under anaerobic condition in order to optimize its genome content and to minimize resources spent by reducing expression of unwanted proteins. Our work also revealed that non-global regulatory transcription factors which control expression of specific regulatory systems are also largely dispensable if there is no selective advantage. For example, in the opportunistic pathogen *P. aeruginosa*, which actively utilizes phenolic compounds, the transcriptional regulators, MhpR, HcaR and FeaR, can sense the compounds and activate target genes that encode enzymes involved in their catabolism. However, more obligate pathogens like *Staphylococcus aureus* and *Campylobacter jejuni*, which do not typically face phenolic compounds in their natural niches, lack both the regulators and their target genes for the utilization of these aromatic compounds. Thus it appears that the absence of any selective pressure to maintain a regulatory protein would render even the global regulatory hubs dispensable, just like any other transcription factor in an organism.

In a recent study on the yeast transcriptional regulatory network [48, 49], we showed that global regulators can be of two distinct types: (i) those that regulate several of their targets in an autonomous manner and (ii) those that integrate signals through different transcription factors and hence combinatorially regulate the expression of their targets. In the light of our finding that there are two classes of global regulatory hubs, we assessed the existence of a trend that global regulators which tend to regulate several targets by themselves would be evolutionarily more conserved than those hubs which co-regulate with several other transcription factors. Our analysis showed that there is no such trend and that both classes of regulatory hubs tend to evolve like any other transcription factor in the genome (fig. 3b). Upon a closer look, we observed that in the *E. coli* transcriptional network, the regulatory hubs which had many target genes were also, in general, the ones which also have many co-regulatory partners indicating that autonomous hubs are far less in number than the integrator-type regulatory hubs.

Given that the global regulatory hubs are not conserved in evolution, we compared the experimentally characterized regulatory network of *E. coli* and *B. subtilis* to understand if there were any differences in the overall topology of the networks from organisms living in different environments. Even though our analysis revealed that the topology of both the networks adopted a similar scale-free structure, it also pointed to the fact that the proteins which emerge as global regulatory hubs in the two organisms are not evolutionarily related. For example, CcpA (which is activated by phosphorylation events) and Crp (which is activated by the presence of cAMP) are the two regulatory hubs in *B. subtilis*

and *E. coli* respectively controlling many genes involved in carbon metabolism. Both have very different modes of regulation and are not evolutionarily related. This suggests that regulatory hubs have been independently innovated to regulate orthologous target genes in organisms living in different environments. These observations provide strong support that the hierarchical structure of these networks has converged to a similar scale-free topology, albeit with independently recruited regulatory hubs. We believe that such an emergence of evolutionarily unrelated proteins to the status of a regulatory hub can be explained because the binding affinity and specificity of a transcription factor and its target site can be affected by relatively small changes in the DNA-binding interface of the transcription factor, or in the binding site [50]. As a result DNA-binding domains could evolve new target sites relatively easily, resulting in rapid de novo emergence of new transcriptional interactions.

Taken together our observations suggest general principles of evolution at the level of global regulatory proteins: (i) transcription factors which are condition specific, be it global regulatory hubs or transcriptional regulators of specific systems, are dispensable in the absence of any selective pressure to maintain them in an organism. (ii) The extent of advantage conferred by orthologous transcription factors to the fitness of an organism might vary across organisms depending on the environment and hence during the course of evolution different proteins may emerge as regulatory hubs in organisms colonizing different niches. (iii) Though different proteins emerge as regulatory hubs, transcriptional regulatory networks tend to approximate a scale-free topology, suggesting that this is a global property, which is enforced entirely independently of the evolutionary forces on the constituent elements of the network [29].

Conclusion

Transcriptional networks, which can be studied at three distinct levels of organization, have been shaped by disparate forces acting at different levels. At the level of the components which comprise these networks, i.e. transcription factors, target genes and regulatory interactions, we observe that (i) the transcription factors complement changes more rapidly than the target genes, with organisms colonizing different ecological niches by evolving their own set of novel transcriptional regulators. This suggests that a major factor in the emergence of new life-styles is the evolution of distinct repertoires of transcription factors, which probably integrate new input signals. (ii) Organisms with similar life-style tend to possess similar regulatory interactions.

In terms of trends which are seen in the evolution of network motifs, we note that (i) network motifs which have the ability to finely regulate the expression

of the target genes are not conserved as rigid units across the different organisms. However, organisms with similar life-style tend to regulate orthologous target genes through similar network motifs suggesting that regulation of genes through appropriate motifs could confer advantage to an organism. (ii) We also note that by losing or conserving specific transcriptional regulators, orthologous genes in different genomes can be incorporated within different regulatory contexts and can thereby easily exhibit different patterns of gene expression. This suggests that natural selection tinkers with individual interactions to arrive at an optimal design to regulate a gene in a given organism.

Finally, at the level of the global network structure, we note that (i) conservation of transcription factors is independent of the number of target genes they regulate or the number of other transcription factors with which a given regulator interacts. Instead, it appears that the determining factor for the retention of a transcriptional regulator appears to be the life-style of the organism. (ii) Additionally, it appears that the same transcription factor can have differential functional relevance for organisms living in different environments due to which evolutionarily unrelated proteins could emerge as hubs in different organisms during the course of evolution. Though different proteins emerge as regulatory hubs, the overall scale-free topology is maintained suggesting that such a structure has evolved convergently and is an emergent property in evolution.

The computational methods developed in our study and those of others, when integrated together with the results from experimental studies which employ recently developed techniques (such as DamID [51], ChIP-chip [52–54], CLIP [55], 1- and 3-hybrid experiments [56] and 2D-EMSA [57]) could complement each other in uncovering the details of transcriptional control in poorly characterized organisms. The predictions from such integrative approaches might allow better design of experiments for biochemical engineering and anti-pathogen therapeutics.

Supplementary URL

Supplementary information detailing our orthology based approach and the predictions of transcriptional networks and transcription factors are available at http://www.mrc-lmb.cam.ac.uk/genomes/madanm/evdy/

Acknowledgements

The authors gratefully acknowledge the Intramural research program of National Institutes of Health, USA for funding their research.

References

1 Jacob F, Monod J: Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 1961;3: 318–356.
2 Ptashne M, Jeffrey A, Johnson AD, Maurer R, Meyer BJ, et al: How the lambda repressor and cro work. Cell 1980;19:1–11.
3 Takeda Y, Ohlendorf DH, Anderson WF, Matthews BW: DNA-binding proteins. Science 1983;221:1020–1026.
4 Pabo CO, Sauer RT: Protein-DNA recognition. Annu Rev Biochem 1984;53:293–321.
5 Browning DF, Busby SJ: The regulation of bacterial transcription initiation. Nat Rev Microbiol 2004;2:57–65.
6 Svetlov VV, Cooper TG: Review: compilation and characteristics of dedicated transcription factors in *Saccharomyces cerevisiae*. Yeast 1995;11:1439–1484.
7 Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, et al: AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis cis*-regulatory elements and transcription factors. BMC Bioinformatics 2003;4:25.
8 Makita Y, Nakao M, Ogasawara N, Nakai K: DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. Nucleic Acids Res 2004; 32(Database issue):D75–D77.
9 Martinez-Bueno M, Molina-Henares AJ, Pareja E, Ramos JL, Tobes R: BacTregulators: a database of transcriptional regulators in bacteria and archaea. Bioinformatics 2004;20:2787–2791.
10 Gonzalez AD, Espinosa V, Vasconcelos AT, Perez-Rueda E, Collado-Vides J: TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes. Nucleic Acids Res 2005;33 (Database issue):D98–D102.
11 Baumbach J, Brinkrolf K, Czaja LF, Rahmann S, Tauch A: CoryneRegNet: an ontology-based data warehouse of corynebacterial transcription factors and regulatory networks. BMC Genomics 2006;7:24.
12 Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al: TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 2006;34 (Database issue):D108–D110.
13 Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, et al: RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res 2006;34(Database issue):D394–D397.
14 Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J: From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. Bioessays 1998;20:433–440.
15 Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA: Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 2004;14:283–291.
16 Barabasi AL, Oltvai ZN: Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5:101–113.
17 Albert R: Scale-free networks in cell biology. J Cell Sci 2005;118:4947–4957.
18 McGuire AM, Hughes JD, Church GM: Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. Genome Res 2000;10:744–757.
19 Tan K, Moreno-Hagelsieb G, Collado-Vides J, Stormo GD: A comparative genomics approach to prediction of new members of regulons. Genome Res 2001;11:566–584.
20 Rajewsky N, Socci ND, Zapotocky M, Siggia ED: The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. Genome Res 2002;12:298–308.
21 Alkema WB, Lenhard B, Wasserman WW: Regulog analysis: detection of conserved regulatory networks across bacteria: application to *Staphylococcus aureus*. Genome Res 2004;14: 1362–1373.
22 Gao F, Foat BC, Bussemaker HJ: Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinformatics 2004;5:31.
23 Rodionov DA, Dubchak I, Arkin A, Alm E, Gelfand MS: Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria. Genome Biol 2004;5:R90.

24 Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, et al: Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res 2004;14:1107–1118.
25 Espinosa V, Gonzalez AD, Vasconcelos AT, Huerta AM, Collado-Vides J: Comparative studies of transcriptional regulation mechanisms in a group of eight gamma-proteobacterial genomes. J Mol Biol 2005;354:184–199.
26 Rodionov DA, Dubchak IL, Arkin AP, Alm EJ, Gelfand MS: Dissimilatory metabolism of nitrogen oxides in bacteria: comparative reconstruction of transcriptional networks. PLoS Comput Biol 2005;1:e55.
27 Barrett CL, Palsson BO: Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach. PLoS Comput Biol 2006;2:e52.
28 Lozada-Chavez I, Janga SC, Collado-Vides J: Bacterial regulatory networks are extremely flexible in evolution. Nucleic Acids Res 2006;34:3434–3445.
29 Madan Babu M, Teichmann SA, Aravind L: Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol 2006;358:614–633.
30 van Nimwegen E: Scaling laws in the functional content of genomes. Trends Genet 2003;19: 479–484.
31 Ranea JA, Buchan DW, Thornton JM, Orengo CA: Evolution of protein superfamilies and bacterial genome size. J Mol Biol 2004;336:871–887.
32 Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM: The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol Rev 2005;29:231–262.
33 Martinez-Antonio A, Janga SC, Salgado H, Collado-Vides J: Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. Trends Microbiol 2006;14:22–27.
34 Kalir S, McClure J, Pabbaraju K, Southward C, Ronen M, et al: Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. Science 2001;292:2080–2083.
35 Mangan S, Alon U: Structure and function of the feed-forward loop network motif. Proc Natl Acad Sci USA 2003;100:11980–11985.
36 Dekel E, Mangan S, Alon U: Environmental selection of the feed-forward loop circuit in gene-regulation networks. Phys Biol 2005;2:81–88.
37 Kalir S, Mangan S, Alon U: A coherent feed-forward loop with a SUM input function prolongs flagella expression in *Escherichia coli*. Mol Syst Biol 2005;1:2005.0006.
38 Mangan S, Itzkovitz S, Zaslaver A, Alon U: The incoherent feed-forward loop accelerates the response-time of the gal system of *Escherichia coli*. J Mol Biol 2006;356:1073–1081.
39 Mayo AE, Setty Y, Shavit S, Zaslaver A, Alon U: Plasticity of the cis-regulatory input function of a gene. PLoS Biol 2006;4:e45.
40 Elena SF, Lenski RE: Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet 2003;4:457–469.
41 Bjornstad ON, Harvill ET: Evolution and emergence of *Bordetella* in humans. Trends Microbiol 2005;13:355–359.
42 Dekel E, Alon U: Optimality and evolutionary tuning of the expression level of a protein. Nature 2005;436:588–592.
43 Fong SS, Joyce AR, Palsson BO: Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. Genome Res 2005;15: 1365–1372.
44 Babu MM, Aravind L: Adaptive evolution by optimizing expression levels in different environments. Trends Microbiol 2006;14:11–14.
45 Conant GC, Wagner A: Convergent evolution of gene circuits. Nat Genet 2003;34:264–266.
46 Madan Babu M, Teichmann SA: Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. Nucleic Acids Res 2003;31:1234–1244.
47 Teichmann SA, Babu MM: Gene regulatory network growth by duplication. Nat Genet 2004;36: 492–496.
48 Balaji S, Babu MM, Iyer LM, Luscombe NM, Aravind L: Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. J Mol Biol 2006;360: 213–227.
49 Balaji S, Iyer LM, Aravind L, Babu MM: Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks. J Mol Biol 2006;360:204–212.

50 Luscombe NM, Thornton JM: Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol 2002;320:991–1009.
51 van Steensel B, Henikoff S: Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nat Biotechnol 2000;18:424–428.
52 Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 2001;409:533–538.
53 Horak CE, Snyder M: ChIP-chip: a genomic approach for identifying transcription factor binding sites. Methods Enzymol 2002;350:469–483.
54 Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al: Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 2002;298:799–804.
55 Ule J, Jensen K, Mele A, Darnell RB: CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods 2005;37:376–386.
56 Drees BL: Progress and variations in two-hybrid and three-hybrid technologies. Curr Opin Chem Biol 1999;3:64–70.
57 Woo AJ, Dods JS, Susanto E, Ulgiati D, Abraham LJ: A proteomics approach for the identification of DNA binding activities observed in the electrophoretic mobility shift assay. Mol Cell Proteomics 2002;1:472–478.

M. Madan Babu and L. Aravind
National Center for Biotechnology Information
National Institutes of Health, Bethesda, MD 20894 (USA)
Tel. +44 1223 402041, Fax +44 1223 213556
E-Mail madanm@mrc-lmb.cam.ac.uk or aravind@ncbi.nlm.nih.gov

Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 81–100

Divergence of Regulatory Sequences in Duplicated Fish Genes

R. Van Hellemont[a], *T. Blomme*[b], *Y. Van de Peer*[b], *K. Marchal*[a,c]

[a]BIOI@SCD, Dept. Electrical Engineering, K.U.Leuven, Heverlee, Leuven, [b]Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Ghent, [c]Dept. Microbial and Molecular Systems, K.U.Leuven, Heverlee, Leuven, Belgium

Abstract

Duplicated genes can undergo different fates, from nonfunctionalization to subfunctionalization and neofunctionalization. In particular, changes in regulatory sequences affecting the expression domain of genes seem to be responsible for the latter two fates. In this study we used in silico motif detection to show how alterations in the composition of regulatory motifs between paralogous genes in zebrafish and *Tetraodon* might reflect the functional divergence of duplicates.

Copyright © 2007 S. Karger AG, Basel

When a gene gets duplicated, it awaits four possible fates. The most likely fate is pseudogenization or nonfunctionalization [1–3]. In rare cases, one of the two duplicates acquires a new function (neofunctionalization; [4]). Subfunctionalization, where both gene copies divide the gene's original functions, forms a third potential fate [5]. Furthermore, recent studies revealed that subfunctionalization is often accompanied by neofunctionalization, which has led to a new model of gene function evolution called sub-neofunctionalization [6]. Finally, both copies can be retained, but, instead of diverging in function, they remain largely redundant and provide the organism with increased genetic robustness against harmful mutations [7, 8]. In addition, retention and redundancy of genes, at least for certain functional classes, is predicted by the 'gene balance' hypothesis, which states that retention of genes with strong dosage effects, such as for instance transcription factors, will be selected against if they are copied without their interacting partners [3, 9, 10].

In particular the subfunctionalization model [2, 5] received much attention of late, since it can, at least partially, explain the large number of genes retained after duplication events, and their subsequent functional divergence [11]. The subfunctionalization model assumes that besides depending on its protein function, the functionality of a gene is also determined by its expression domain (where and when the gene is expressed). The specific expression domain of a gene results at least partially from its transcriptional regulation, which is, in turn, encoded by a specific combination of transcription factor binding sites (defined as a regulatory module) in the gene's promoter. Each transcription factor binding site (TFBS) in a module corresponds to a DNA consensus sequence or motif that is recognized by its cognate regulatory protein or transcription factor. Changes in these TFBS can therefore be an important antecedent for expression divergence and thus for sub- or neofunctionalization [5, 8, 12–18]. However, the number of studies that show how expression divergence between paralogs is reflected by differences between regulatory elements of paralogous gene pairs is still limited [19, 20].

As a result of a genome wide fish-specific duplication event that occurred some 350 mya [21–23] and of more recent duplication events [9], ray-finned fish such as *Tetraodon* and zebrafish contain a large number of duplicated genes [24–27], of which several have already been shown to have undergone subfunctionalization [25, 28–33]. In this study, we further investigate to what extent 'in silico' analyses support expression divergence of genes through identified changes in regulatory sequences. To this end, we explicitly searched for motifs that have been preserved over 450 mya of vertebrate evolution (from mammals to ray-finned fish), but have been differentially retained in either one of two duplicates in zebrafish or *Tetraodon* and that thus might explain the experimentally observed expression divergence.

Methodology

Identification of Suitable Data Sets

In this study we focused on duplicated fish genes for which there was (some) experimental evidence that supported subfunctionalization, such as for *bmp2* [34], *glyRα* [35], *msx2* [36], *pax6* [37, 38], and *shh* [39]. In addition, several other duplicated fish genes were included, namely *efna1, en2* [5, 40], *kcnip1, ntng1, ntng2* and *six4*, and a *glyRα*-related gene family. Phylogenetic trees were constructed based on the predicted protein sequences (Ensembl [41] release 37) from human, mouse, *Tetraodon nigroviridis*, zebrafish (*Danio rerio*), rat (*Rattus norvegicus*), chicken (*Gallus gallus*), and frog (*Xenopus tropicalis*). If splice variants were reported, the longest transcript was used.

To delineate vertebrate gene families, a similarity search was performed (BLASTP, [42]; E-value cutoff E-10) with all proteins from the organisms listed above, plus those of *Ciona intestinalis* [43], version 1 and *Drosophila melanogaster* (Ensembl [41]), version 3, which were added as outgroup species. Blast hits between vertebrate sequences with a score better than the best score between a vertebrate sequence and an outgroup sequence (*Drosophila* or *Ciona*) were retained and considered members of the same gene family. The *Drosophila* or *Ciona* sequence was used to root the phylogenetic tree (see further). For each gene family, a multiple sequence alignment was created with T-Coffee 1.37 using default parameters [44]. Alignment columns containing gaps were removed when a gap was present in more than 10% of the sequences. To reduce the chance of including misaligned amino acids, all positions in the alignment left or right from the gap were also removed until a column in the sequence alignment was found where the residues were conserved in all genes included in our analyses. This was determined as follows: for every pair of residues in the column, the BLOSUM62 value was retrieved. If at least half of the pairs had a BLOSUM62 value ≥ 0, the column was considered as conserved. Neighbor joining trees (with 500 bootstrap replicates) were constructed with PHYLIP 3.5 [45] using both nucleic and amino acid sequence alignments and simple Poisson-corrected substitution models.

For all datasets, the phylogenetic tree showed genes duplicated in at least one of the fish species. These duplicates, and their homologs in human, chicken and frog were selected for further analysis (table 1). Intergenic regions of these homologs were retrieved using the Ensembl mart database release 37. Intergenic regions were defined as the region upstream of the transcription start (as defined by Ensembl), limited to 2 kb and including the 5′UTR.

Pairwise Alignment of Paralogous Intergenic Sequences

Pairwise alignments of paralogous intergenic regions were obtained with Smith-Waterman using the default parameters (gap open penalty 10 and gap extension penalty 0.5) [46]. The expected percentage identity between two unrelated intergenic sequences of the same organism was estimated by averaging the scores obtained by aligning each intergenic sequence against all other intergenic sequences of the same organism but not belonging to the same protein family.

Search for Regulatory Motifs Conserved in Each of the Gene Sets

For each gene set (i.e., all genes belonging to the same gene family), intergenic sequences were subjected to BlockSampler [47]. BlockSampler was run using default parameters – searching plus strand only ($s = 0$) and searching for

Table 1. Description of the datasets

Dataset	Newick tree[a]	Ensembl Gene IDs[a]	Experimental evidence
*bmp2**	((Xt, (Hs, Gg)), (Dr1, (Dr2, Tn)));	(Dr1) ENSDARG00000013409, (Dr2) ENSDARG00000041430, (Gg) ENSGALG00000008830, (Hs) ENSG00000125845, (Tn) GSTENG00020275001, (Xt) ENSXETG00000005519	RT-PCR + in situ hybridization [34]
Efna1	(Hs, ((Dr1, Tn1), (Dr2, Tn2)));	(Dr1) ENSDARG00000030326, (Dr2) ENSDARG00000018787, (Hs) ENSG00000169242, (Tn1) GSTENG00032578001, (Tn2) GSTENG00033951001	/
en2	((Xt, Hs), (Dr1, (Dr2, Tn)));	(Dr1) ENSDARG00000026599, (Dr2) ENSDARG00000038868, (Hs) ENSG00000164778, (Tn) GSTENG00023985001, (Xt) ENSXETG00000013496	/
*glyRα1**	(Gg, ((Dr, Tn1), Tn2)));	(Dr) ENSDARG00000006865, (Gg) ENSGALG00000004936, (Tn1) GSTENG00029286001, (Tn2) GSTENG00022245001	in situ hybridization [35]
glyRα1-related	((Xt, Gg), (Dr1, (Dr2, Tn)));	(Dr1) ENSDARG00000012019, (Dr2) ENSDARG00000011066, (Gg) ENSGALG00000004134, (Tn) GSTENG00024269001, (Xt) ENSXETG00000001966	/
kcnip1	((Xt, (Gg, Hs)), ((Dr1, Tn1), (Dr2, Tn2)));	(Dr1) ENSDARG00000034808, (Dr2) ENSDARG00000022109, (Gg) ENSGALG00000002132, (Hs) ENSG00000182132, (Tn1) GSTENG00020358001, (Tn2) GSTENG00024581001, (Xt) ENSXETG00000018293	/
*msx2**	((Xt, (Gg, Hs)), (Dr1, Dr2));	(Dr1) ENSDARG00000009936, (Dr2) ENSDARG00000006982, (Gg) ENSGALG00000002947, (Hs) ENSG00000120149, (Xt) ENSXETG00000009168	in situ hybridization [36]
ntng1	((Gg, Hs), (Tn1, (Dr, Tn2)));	(Dr) ENSDARG00000014973, (Gg) ENSGALG00000001896, (Hs) ENSG00000162631, (Tn1) GSTENG00027711001, (Tn2) GSTENG00035109001	/

ntng2	((Hs, Gg), ((Dr, Tn1), Tn2));	(Dr) ENSDARG00000036938, (Gg) ENSGALG00000003677, (Hs) ENSG00000196358, (Tn1) GSTENG00004089001, (Tn2) GSTENG00014392001	/
*pax6**	(((Gg, Hs), Xt), ((Dr1, Dr2), Tn));	(Dr1) ENSDARG00000045045, (Dr2) ENSDARG00000045936, (Gg) ENSGALG00000012123, (Hs) ENSG00000007372, (Tn) GSTENG00025814001, (Xt) ENSXETG00000008175	in situ hybridization + transient transfection assays + western blot analysis [37]
*shh**	((Hs, Gg), (Dr1, (Dr2, Tn)));	(Dr1) ENSDARG00000038867, (Dr2) ENSDARG00000039710, (Gg) ENSGALG00000006379, (Hs) ENSG00000164690, (Tn) GSTENG00023991001	in situ hybridization [39]
six4	((Xt, Hs), ((Dr1, Tn), Dr2));	(Dr1) ENSDARG00000031983, (Dr2) ENSDARG00000004695, (Hs) ENSG00000100625, (Tn) GSTENG00032223001, (Xt) ENSXETG00000016941	/

Dataset: Indicates the name of the gene family (derived from the human ortholog in the dataset). For gene sets indicated with an asterisk, experimental evidence supporting expression divergence between the fish paralogs exists. Newick tree: for each dataset the phylogenetic relations are given in Newick format. Ensembl Gene IDs: lists the genes present in each dataset by their ensembl gene ID. Experimental evidence: indicates the type of experimental evidence that supports expression divergence.

[a]Dr: *Danio rerio*, Gg: *Gallus gallus*, Hs: *Homo sapiens*, Tn: *Tetraodon nigroviridis*, Xt: *Xenopus tropicalis.*

one motif per run, prior set to 0.2, initial motif length of 8 nt, and a threshold on consensus score of 1.0. BlockSampler requires the definition of a root sequence, i.e. only conserved motifs, which are also present in the root, will be retained. As for our application the biological meaning of a root was less clear, each sequence of the gene set was chosen once as root. Per root sequence BlockSampler was run 100 times implying that the total number of runs and retrieved motifs for a gene set equaled 100 times the 'number of sequences in the gene set'. The motifs with a consensus score above 1 were selected and motifs overlapping for more than 80% were merged to avoid redundancy.

In order to account for the fact that short motifs are more likely to have a higher degree of conservation than long motifs, the consensus score of each detected block was normalized for the length of the motif using the following formula, $Cs_{ad} = (L/(L + E))\ Cs$, where L is the length of the conserved block, E is an empirical factor (set to 5) and Cs the consensus score [47].

Assessing the Statistical Significance of Detected Motifs

For each gene set, 30 random sets were compiled. These random sets have a composition similar to the genuine gene set in sequence number and origin (species), but in contrast to the genuine gene set sequences were selected randomly and as a result do not share any homology relation. For each random set, we performed the same analysis as for the genuine gene sets: BlockSampler was applied to identify conserved motifs. Per random set, the number of runs equaled 100 times the number of sequences in the random set (of which each one served once as root). After normalizing the scores, from the 100 runs of a single root the best scoring motif (highest Cs_{ad}) was selected. This resulted, for each genuine gene set in a number of random motifs equaling 30 (i.e., the number of random sets) times the 'number of sequences in this random set (i.e., number of root sequences)'. The scores of these motifs were used to estimate a random motif score distribution. To identify significant motifs in the genuine dataset we chose the Cs_{ad} of the xth percentile of this random distribution as a threshold. As a result, motifs in the genuine dataset with a Cs_{ad} higher than the chosen threshold were considered statistically significant.

Identifying Motifs Supporting Subfunctionalization

Motifs that potentially support the subfunctionalization model were identified using the following criteria: a motif was considered if it was conserved over a region of at least 8 nt and lost in at least one paralog of the fish species for which multiple paralogs were present in the gene set. In order to minimize false positive motifs, extra constraints were set on the number of additional species in which the motif had to be conserved. Indeed, if conserved over larger

Table 2. Motifs indicative for subfunctionalization with a Cs_{ad} score exceeding the 90th percentile of the random score distribution

Dataset	#	L	PI	Conservation profile	Subf sp.	Duplication type	Motif name
*bmp2**	1	29	99.5	Dr2_Gg_Hs_Tn	Dr	FSD	bmp2_1
*pax6**	4	70	90	Dr2_Gg_Hs_Tn_Xt	Dr	ZS	pax6_1
		38	90	Dr2_Gg_Hs_Tn_Xt	Dr	ZS	pax6_2
		71	90	Dr2_Gg_Hs_Tn_Xt	Dr	ZS	pax6_3
		50	90	Dr2_Gg_Hs_Tn_Xt	Dr	ZS	pax6_4
*shh**	2	15	95	Dr2_Gg_Hs_Tn	Dr	FSD	shh_1
		16	99	Dr2_Gg_Hs_Tn	Dr	FSD	shh_2
kcnip1	1	13	95	Dr1_Dr2_Gg_Hs_Tn1_Xt	Tn	FSD	Kcnip1_1
Total	8						

Dataset: The gene set in which the motifs were detected. Gene sets indicated with an asterisk contain fish paralogs for which subfunctionalization has been shown in literature. #: the number of motifs detected in the gene set that support subfunctionalization given this threshold. L: the length of the motif indicative for subfunctionalization (number of nucleotides) PI: indicates the percentile of the random distribution to which the score of the motif belongs. Conservation profile: indicates in which homologs of the gene family the motif was also present (see footnote of table 1). Subf sp.: indicates in which fish-species the motif was lost. Duplication type: indicates from which duplication event the paralogs originated for which a motif was found that supports subfunctionalization (FSD: ancient fish specific duplication event; TS: *Tetraodon* specific duplication event; ZS: zebrafish specific duplication event). Motif name: the name to unambiguously indicate a specific motif.

phylogenetic distances, we can be more confident in the motif prediction. These constraints depended on the composition of the gene set: If the gene set under study consisted of multiple non-fish homologs, either frog, chicken or human (which was the case for *bmp2*, *en2*, *glyRα*-related, *kcnip1*, *msx2*, *ntng1*, *ntng2*, *pax6*, *shh*, *six4*), a motif was only considered if it was conserved in at least two non-fish species. If the motif under study was derived from a gene set that contained only one non-fish homolog (*efna1* and *glyRα*), the motif had to be present in this one non-fish sequence.

For each motif we constructed a profile that indicates whether or not the motif occurs in the respective species from which the homologs of the gene family were derived. If the profile of the motif satisfies the requirements mentioned above, its profile was said to support subfunctionalization. Phylogenetic profiles of motifs supporting subfunctionalization are represented in tables 2 and 3. To assess whether the detected motifs correspond to known transcription factor binding sites, we scanned the human instance of each conserved motif with the Transfac 8.2 database of vertebrate transcription factor binding site profiles [48]. This scanning was performed using MotifLocator [49, 50] with a

Table 3. Motifs indicative for subfunctionalization for the gene sets for which a relaxed selection criterium was used

Dataset	#	L	PI	Conservation profile	Subf sp.	Duplication type	Motif name
efna1	2	12	75	Dr1_Dr2_Hs_Tn2	Tn	TS	efna1_1
		8	55	Dr1_Dr2_Hs_Tn2	Tn	TS	efna1_2
Total	2						

For column descriptions and legend see tables 1 and 2.

0th order vertebrate background model. Hits with a score >0.9 were regarded as potential binding sites. The binding sites are indicated by the Transfac factor name [48]. To further validate the link between the binding sites revealed with this screening and the gene under study, we did a text-based search with PubMed [51] using the name of the gene/protein under study (e.g. *pax6*) and the name of the transcription factor potentially binding the promoter region of this gene as search terms. When such a link existed, this is explicitly mentioned in the results section.

Results

The goal of our study was to see whether divergent expression of duplicated genes is reflected in any detectable way by a different composition of their regulatory sequences, in particular by the presence or absence of specific motifs. First, this requires identifying interesting case studies, i.e., gene families that contain members of fish specific duplication events. Second, we need to compile the potential regulatory motifs present in the intergenic sequences of these gene families and to identify which of the motifs have been differentially retained in one of the fish paralogs. However, since the regulatory motifs present in fish genomes are still largely unknown, the list of potential motifs was compiled based on comparative de novo motif detection methods, better known as phylogenetic footprinting. Phylogenetic footprinting assumes that biologically relevant sequences, such as regulatory motifs, evolve slower than their surrounding non-functional intergenic sequences. By using cross-species conservation, short stretches of DNA that are conserved over certain phylogenetic distances are identified as potential motifs. The greater the phylogenetic distance over which the motif is conserved and the more orthologs in which the

motif can be detected are present, the more confidence can be put in this prediction.

However, as we are specifically searching for conserved motifs that are differentially lost between paralogs, we had to rely on a phylogenetic footprinting methodology that is able to align strongly evolved sequences of which some do not contain the motifs [47].

Identifying Gene Sets Containing Duplicated Fish Genes

Our analysis was performed on a selection of gene families that contained paralogs either originating from a duplication event before the divergence of zebrafish and *Tetraodon* (further referred to as the ancient fish specific duplication, FSD) or from a more recent duplication specific to either zebrafish or *Tetraodon* (that occurred after divergence of both species [9]) (for a complete list of these gene families see table 1). For some of these fish-specific paralogs, subfunctionalization was supported by literature (e.g., *bmp2* [34, 39], *glyRα* [35], *msx2* [36], *pax6* [37, 38] and *ssh* [39] (see table 1)).

For gene sets *bmp2*, *efna1*, *en2*, *glyRα1*, *glyRα1*-related, *kcnip1*, *ntng1*, *ntng2*, *pax6*, *shh* and *six4*, the topology of the corresponding phylogenetic trees indicates that the paralogs resulted from the ancient FSD event which took place before the divergence of zebrafish and *Tetraodon* (about 150 mya [24]). For the *pax6* gene family, the two zebrafish copies are the result of a more recent zebrafish specific duplication event (table 1). Concerning the *msx2* gene family, the topology did not allow us to conclude whether the *msx2* zebrafish copies resulted from the ancient FSD or whether they were the result of a more recent duplication event in zebrafish.

Determining the Overall Homology Between Paralogous Intergenic Regions

In order to determine their overall conservation, intergenic paralogous regions in fish were aligned using Smith-Waterman [46]. Results are shown in table 4. The conservation level of paralogous intergenic regions resulting from the ancient FSD [9] (*Tetraodon*: 40.4% and zebrafish: 43.6%) was comparable to that of unrelated sequences, which was estimated 40.4% and 43% for *Tetraodon* and *Danio rerio* respectively (see methods section). This analysis also indicates that in these ancient duplicates, except for the conserved regulatory motifs, no sequence conservation is to be expected.

Identification of Motifs Supporting Subfunctionalization

To search for differences in motif composition between duplicates, we first compiled all potential motifs conserved within the intergenic regions of genes belonging to the same gene family (see material and methods). To this end we

Table 4. Intergenic homology between fish duplicates

Dataset	Compared duplicates	Type	Identity (%)
*bmp2**	Dr1–Dr2	FSD	44.4
*glra**	Tn1–Tn2	FSD	41.4
*msx2**	Dr1–Dr2	FSD/ZS	44.5
*pax6**	Dr1–Dr2	ZS	39.2
*shh**	Dr1–Dr2	FSD	43.6
efna1	Dr1–Dr2	FSD	43.9
	Tn1–Tn2	FSD	32.6
en2	Dr1–Dr2	FSD	45.9
glra-related	Dr1–Dr2	FSD	42.3
kcnip1	Dr1–Dr2	FSD	43.4
	Tn1–Tn2	FSD	41.9
ntng1	Tn1–Tn2	FSD	42.3
ntng2	Tn1–Tn2	FSD	43.9
six4	Dr1–Dr2	FSD	41.7

Intergenic regions of fish paralogs present in the gene sets under study were pairwise aligned using Smith-Waterman [46]. Dataset: indicates the name of the gene family (derived from the human ortholog in the dataset) from which the fish paralogs were compared. For gene sets indicated with an asterisk, experimental evidence supporting expression divergence between the fish paralogs exists. Compared duplicates: the fish genes for which the intergenic sequences were aligned; for the corresponding Ensembl gene IDs and legend we refer to table 1. Type: the type of duplication event the duplicates are the result of (based on the phylogenetic trees); FSD: ancient fish specific duplication; ZS: zebrafish specific duplication. Percent identity: the similarity between the intergenic sequences of the aligned duplicates.

used BlockSampler, a procedure based on Gibbs sampling that searches for statistically overrepresented motifs. In the presence of an appropriate background model, the procedure is known to be quite robust against noise, i.e., sequences that do not contain the motif [52–54]. In the context of subfunctionalization, this property is essential as it allows finding motifs that are not conserved in all branches of the phylogenetic tree. For each set, applying BlockSampler resulted in a list of conserved motifs. To detect motifs supporting subfunctionalization, from this list we selected those motifs that were significantly conserved but missing in at least one of the fish paralogs, either *Tetraodon* or zebrafish (i.e., the species for which multiple paralogs are present in the gene set under study). Especially for the ancient duplications for which the overall similarity in intergenic sequences between the fish paralogs is quite low, many differences are

expected to be found in their promoter regions, most of which probably do not correspond to biologically relevant subfunctionalized motifs. Therefore, in order to select the most relevant predictions we considered only those motifs that were also conserved in phylogenetic lineages other than fish and thus were preserved over 450 mya of vertebrate evolution (see materials and methods for the exact criteria).

To test to what extent the choice of the threshold on the motif scores (defined as the xth percentile of the random scores) determined the total number of motifs retrieved and thus the number of gene sets for which we detected a motif(s) indicative for subfunctionalization, we repeated the analysis for multiple threshold levels (ranging from the 99.5th to the 50th percentile of the random score distribution). The results for gene sets containing homologs from multiple non-fish species are summarized in table 2, considering the 99.5th, 99th, 95th and 90th percentile of the random distribution as threshold on the motif scores. As expected, lowering the threshold of our search allows detecting more motifs indicative for subfunctionalization. However, as the stringency of the search becomes lower, the motifs taken into account become gradually shorter and presumably less reliable.

When using a quite conservative threshold (motif scores exceeding the 90th percentile of the random distribution), in three (*bmp2*, *pax6* and *shh*) out of the five datasets for which expression divergence was experimentally demonstrated, we could find at least one motif indicative for subfunctionalization. Besides in these experimentally supported datasets, we also found motif-based indications for subfunctionalization in *efna1* (although only when using a relaxed threshold in the motif scores, table 3) and *kcnip1*.

Detailed Description of the Datasets with Subfunctionalized Motifs

Figures 1 to 5 display the results for the datasets *bmp2*, *pax6*, *shh*, *kcnip1* and *efna1*. Significantly overrepresented motifs are mapped. An arrow indicates motifs that might be supportive of subfunctionalization. Below we give a more detailed description of these results.

In vertebrates, bone morphogenetic proteins **(Bmps)** play a crucial role in establishing the early body plan and in organogenesis [55]. Martinez-Barbera et al. [34] studied the expression pattern of zebrafish *bmp2* paralogs, *bmp2a* and *bmp2b*. They found indications for divergent expression profiles in the gastrulating embryo and in the pectoral fin bud. In this study, the *bmp2* gene family consists of two zebrafish genes (table 1, fig. 1) that correspond to the genes studied by Martinez-Barbera et al. [34]. The motif indicated in figure 1 (table 2) that has been retained in one zebrafish copy (Dr2, ENSDARG00000041430) but that was lost in the other (Dr1, ENSDARG00000013409), could possibly explain this observed divergence.

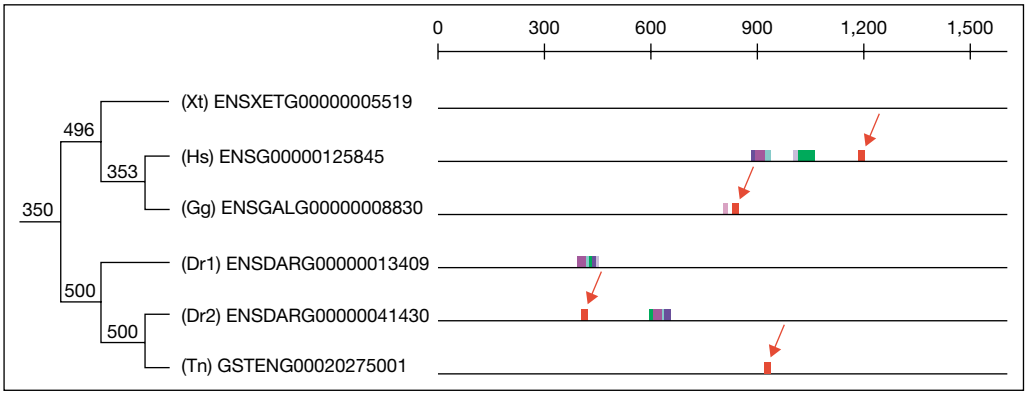


Fig. 1. Graphical display of the motifs found in the upstream regions of *bmp2*. Graphical display of all motifs with a score exceeding the 90th percentile of the random score distribution. The motifs that are in support of subfunctionalization are indicated by an arrow. The phylogenetic tree (branch lengths not drawn to scale) illustrates the evolutionary relationships between the homologs; these are indicated as defined in table 1. Abbreviations used: Xt: *Xenopus tropicalis*, Hs: *Homo sapiens*, Gg: *Gallus gallus*, Dr: *Danio rerio*, Tn: *Tetraodon nigroviridis.*

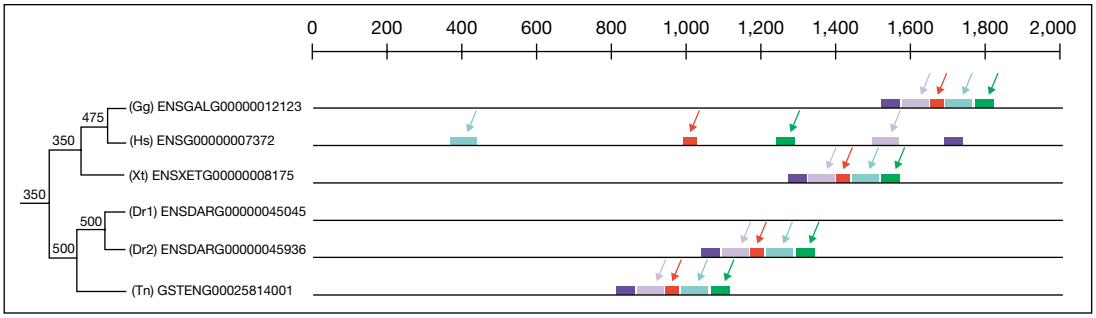


Fig. 2. Graphical display of the motifs found in the upstream regions of *pax6.* Interpretation is as in figure 1.

Pax6 plays an important role in the central nervous system and in the developing eye of both vertebrates and invertebrates. According to our analysis, the *pax6* gene family contains two zebrafish paralogs, which (given the position of the *Tetraodon* homolog in the tree topology) originated from a zebrafish specific duplication (fig. 2). The presence of two zebrafish paralogs is consistent with the observations of Nornes et al. [37]. They observed that both zebrafish

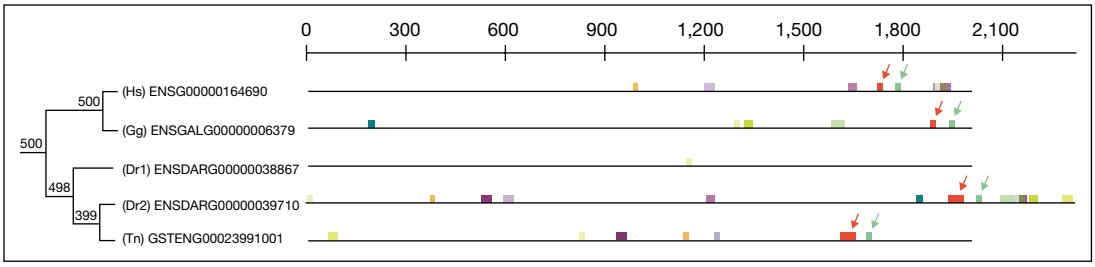


Fig. 3. Graphical display of the motifs found in the upstream regions of *shh.* Interpretation is as in figure 1.

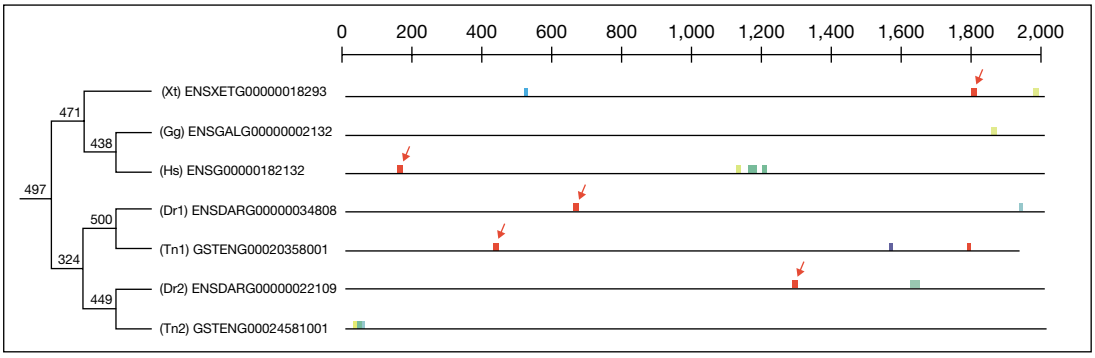


Fig. 4. Graphical display of the motifs found in the upstream regions of *kcnip1.* Interpretation is as in figure 1.

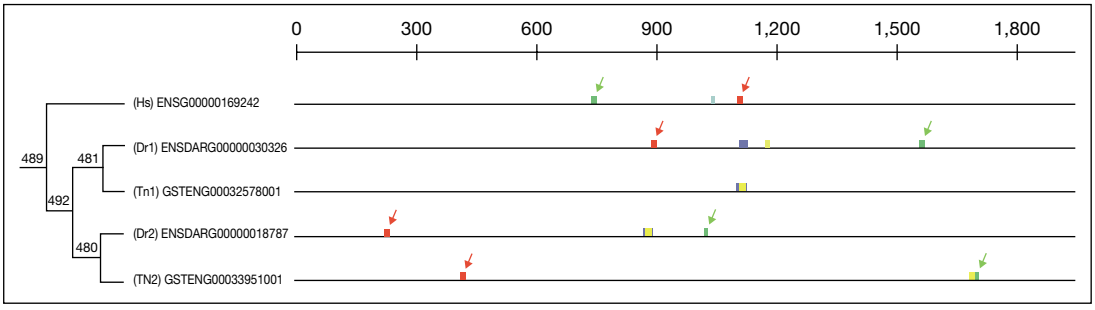


Fig. 5. Graphical display of the motifs found in the upstream regions of *efna1.* Interpretation is as in figure 1.

copies have unique expression domains that sum up to the total expression domain for the single *pax6* copy present in birds and mammals [2]. Figure 2 displays the conserved motifs identified in the promoter region of the *pax6* homologs. Motifs that might be indicative for subfunctionalization are indicated by arrows: we identified four motifs conserved in human, chicken, frog, *Tetraodon* and in one zebrafish paralog (Dr2, ENSDARG00000045936; table 2). The complete absence of all of these motifs in the zebrafish paralog (Dr1, ENSDARG00000045045) can also be interpreted as an indication for nonfunctionalization of this paralog. However, because experimental evidence about the expression of both zebrafish paralogs exists [37], subfunctionalization seems the more likely fate. The order in which the motifs occur in the intergenic regions seems to be perfectly conserved in the non-mammalian sequences where they are concatenated into a large conserved region of circa 250 nt (fig. 2). In the human ortholog on the contrary, the order and spacing of these motifs seems to be altered. In order to get an idea of the binding sites localized in the four motifs reported here, we screened the human motif instance with the Transfac database of transcription factor binding sites. As is summarized in table 5, different potential binding sites are present in the *pax6* motifs. For instance, pax6_3 contains an AP-2α and an AP-2rep binding site. This is plausible, since both Pax6 and AP-2α function in eye development [56]. Moreover, both transcription factors are known to interact in coordinating corneal epithelial repair [57].

Vertebrate **hedgehog** genes are involved in many developmental processes [58]. As Laforest et al. showed that zebrafish hedgehog paralogs exhibit expression patterns that suggest subfunctionalization, we chose to study the *sonic hedgehog* or *shh* gene family [39] in more detail. Figure 3 illustrates two significant motifs (see arrows) that possibly support subfunctionalization (see also table 2). These motifs, indicated in red and green, have both been conserved in human, chicken, *Tetraodon* and Dr2 (ENSDARG00000039710) but were lost in Dr1 (ENSDARG00000038867). The order and spacing between these two motifs also seems to have been retained during evolution. Besides these, figure 3 also displays some additional interesting motifs pointing towards subfunctionalization (for instance, the dark purple and dark yellow motifs). These were not initially retained as 'significant motifs' under our strong selection criteria, because they were either too short or not conserved in multiple non-fish species.

kcnip1 encodes the potassium channel-interacting protein [59]. In this study we identified a frog, chicken and human homolog, two zebrafish and two *Tetraodon* paralogs. The tree topology (fig. 4) indicates that the four fish genes are the result of an ancient FSD. As is shown in figure 4, we identified one motif that is in support of a possible divergent expression profile (table 2). This

Table 5. The potential transcription factor binding sites located in the detected motifs indicative for subfunctionalization

Motif Name	Consensus and possible binding sites
bmp2_1	TTGTTTTGTTTTGTTTTTT SRY, M00148: AAACWAM: 5-11 – (1.0); 10-16 – (1.0)
pax6_1	GGCTCGAGGGCCAGGTTGAGGGTACTCATCGAGCCTCGAACTCCTCCTAAAAATGATTCCTGCCAAAAGC Cap, M00253, NCANHNNN: 49-56 – (0.963) CdxA, M00101, AWTWMTR: 46-52 – (0.904) Hnf4, M00967, AARGTCCAN: 6-14 + (0.931) Etf, M00695, GVGGMGG: 43-49 – (0.906) Lyf-1, M00141, TTTGGGAGR: 59-67 – (0.931); 43-51 – (0.950) NF1, M00193, NNTTGGCNNNNNNCCNNN: 51-68 – (0.919)
pax6_2	ACCACTGTCACTTTCAAATTGGAGAGCCAGATGGAAGC E2a, M00804, GGCGSG: 21-34 – (0.907) Irf, M00772, BNCRSTTTCANTTYY: 1-15 + (0.946) Tal1, M00993, TCCAKCTGNY: 26-35 – (1.0)
pax6_3	TGGTAAGGTCTAGGCCCAGACTAGAGTGGCCAGTGGGAGGTGGGCGCTCCTAGGCCTTAACACAGGATGCC AP-2α, M00469, GCCNNNRGS: 29-37 + (0.917) AP-2rep, M00468, CAGTGGG: 31-37 + (1.0) Cap, M00253, NCANHNNN: 16-23 + (0.906); 63-70 – (0.925); 36-43 – (0.904) CCAAT box, M00254, NNNRRCCAATSA: 25-36 + (0.933) C/EBP, M00159, NNTKTGGWNANNN: 54-66 – (0.906) CHCH, M00986, CGGGNN: 34-39 + (0.932) Etf, M00695, GVGGMGG: 34-40 + (0.911) Ets, M00971, ACTTCCTS: 63-70 – (0.925) Pea3, M00655, ACWTCCK: 64-70 – (0.933)

Table 5. (continued)

Motif Name	Consensus and possible binding sites
pax6_4	ATTTTCCTGTTTTCCTCCTCTAAGTCACAAAGTCAACAGTTAATTCAAAG AP-1, M00172, RSTGACTNMNW: 19-29 – (0.942) AP-1, M00517, NNNTGAGTCAKCN: 18-30 – (0.905) AP-1, M00924, TGACTCANNSKN: 16-27 – (0.901) AP-1, M00926, TGAGTCAN: 21-28 + (0.904) Fox, M00809, KATTGTTTRTTTW: 29-41 – (0.953) Hnf3α, M00724, TRTTTGYTYWN: 28-38 – (0.932) Hnf3β, M00131, KGNANTRTTTRYTTW: 29-43 – (0.908) Pou1f1, M00744, ATGAATAAWT: 39-48 – (0.915) Sf1, M00727, TGRCCTTG: 28-35 – (0.918) Stat1, M00496, NNTTTCCN: 1-8 + (0.948); 9-16 + (0.9704) Stat6, M00500, NNYTTCCY: 9-16 + (0.915)
shh_1	GCTCTCCAGGCTTGC
shh_2	TCAGATGCGCCCCTGG
kcnip1_1	TGTGTATCTGTGT
efna1_1	ACGCAGACACACA
efna1_2	ATGTTTATT

Motif name: The name to unambiguously indicate a certain motif detected (with a Cs_{ad} score exceeding 90th percentile of the random score distribution for *bmp2*, *pax6*, *ssh* and *kcnip1* and a Cs_{ad} score exceeding 50th percentile for *efna1*). These names correspond to the ones in tables 2 and 3. Consensus and possible binding sites: the sequence of the motif in the intergenic region of the human homolog (table 1) is given followed by the possible binding sites situated in this motif (Transfac name, Transfac ID, consensus sequence, positions, strand and score). Remark: For the shortest motifs MotifLocator could not be used to screen for potential binding sites. Therefore only the motif instance in human is given.

motif, indicated in red, is retained in frog, human, both zebrafish paralogs, and one *Tetraodon* **paralog** (Tn1, GSTENG00020358001). Two other interesting motifs (the green and light blue motifs) were detected in Tn2, GSTENG00 024581001: both these motifs seem to be present in the human sequence but are divergently retained over the fish paralogs. The smaller blue motif is retained in Dr1 and Tn1, while the green motif is retained in Dr2 and Tn2. From the tree topology it seems that the combined motif, still present in the human sequence might have been subfunctionalized after an early fish duplication that took place before the speciation between *Tetraodon* and zebrafish. The motifs seem to have a classical pattern of subfunctionalization. Note, however, that we did not primarily retain them as they do not meet our selection criteria (the motifs are conserved in one non-fish homolog only).

Also in ***efna1***, which encodes an ephrin-A1 precursor, we found two motifs indicative of expression divergence between paralog Tn1 (GSTEN G00032578001) and paralog Tn2 (GSTENG00033951001; fig. 5). These two motifs, respectively 12 and 8 bp long, and conserved in the intergenic sequences of the human homolog were also present in both zebrafish paralogs, but only in one of the two *Tetraodon* paralogs (Tn2; see table 3).

Discussion

In this study, we found indications that expression divergence between paralogs in zebrafish and/or *Tetraodon* is reflected by differences in regulatory motifs. We investigated five gene families for which experimental evidence supported subfunctionalization. For three of these proof-of-concept gene families, we identified at least one motif that was differentially lost after a fish-specific duplication event and that seemed to be in accordance with the experimentally observed expression divergence. Besides in the 'proof-of-concept' datasets, we found differential alterations in regulatory motifs between the fish paralogs that point towards potential subfunctionalization in two other gene families (*efna1* and *kcnip1*).

In order to assess which potential transcription factors bind to the conserved motifs, we screened them with the Transfac database of TFBS. Several potential TFBS seemed to be present in the conserved motifs but to our knowledge, for the majority of these TFBS no clear link with the genes containing the conserved motifs was found in literature.

The sequence dependent indications for subfunctionalization identified in this study of course largely depend on the reliability of our in silico predicted regulatory motifs. To select confident predictions, we used strict selection criteria and considered only those motifs that were conserved over at least 450 mya of vertebrate

evolution. On the other hand, by using these conservative selection criteria we probably discard many functional motifs. Indeed, motifs that are too short, too degenerated, or very lineage specific will remain undetected. As a result, we most likely underestimate the number of motifs indicative for subfunctionalization. This might explain why we only find in a subset of the 'proof-of-concept' datasets sequence based indications for subfunctionalization. Moreover, according to its strict definition, subfunctionalization implies that both paralogs divide the gene's original function over both gene copies. When relating this to expression divergence and subsequent changes in regulatory motifs, one expects to find two ancestral motifs still present in an outgroup species to be divided between the two paralogous intergenic regions. We could not detect any example of this idealized situation of subfunctionalization due to our conservative approach; however, when using more relaxed criteria our method identified such an example.

Despite our conservative strategy, in nearly half of the tested datasets clear sequence based indications for potential expression divergence were present, indicating that subfunctionalization is probably more general than is assumed at this point (see also [8]).

Acknowledgements

T. Blomme and R. Van Hellemont are fellows of the IWT. This work is partially supported by: 1. IWT projects: GBOU-SQUAD-20160; 2. Research Council KULeuven: GOA Mefisto-666, GOA-Ambiorics, EF/05/007 SymBioSys, IDO genetic networks; 3. FWO projects: G.0115.01, G.0413.03 and G.0318.05; 4. IUAP V-22 (2002–2006). We would like to thank S. Robbens for the assistance with the construction of phylogenetic trees.

References

1 Lynch M, Conery JS: The evolutionary fate and consequences of duplicate genes. Science 2000;290: 1151–1155.
2 Lynch M, Force A: The probability of duplicate gene preservation by subfunctionalization. Genetics 2000;154:459–473.
3 Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, et al: Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA 2005;102:5454–5459.
4 Taylor JS, Raes J: Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet 2004;38:615–643.
5 Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: Preservation of duplicate genes by complementary, degenerative mutations. Genetics 1999;151:1531–1545.
6 He X, Zhang J: Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 2005;169:1157–1164.
7 Gu X: Evolution of duplicate genes versus genetic robustness against null mutations. Trends Genet 2003;19:354–356.
8 Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y: Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. Genome Biol 2006;7:R13.1–R13.11.

9 Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y: The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol 2006;7:R43.1–R43.12.
10 Freeling M, Thomas BC: Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res 2006;16:805–814.
11 Moore RC, Purugganan MD: The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol 2005;8:122–128.
12 Papp B, Pal C, Hurst LD: Evolution of *cis*-regulatory elements in duplicated genes of yeast. Trends Genet 2003;19:417–422.
13 De Bodt S, Theissen G, Van de Peer Y: Promoter analysis of MADS-box genes in eudicots through phylogenetic footprinting. Mol Biol Evol 2006;23:1293–1303.
14 Prince VE, Pickett FB: Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet 2002;3:827–837.
15 Gu X, Zhang Z, Huang W: Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc Natl Acad Sci USA 2005;102:707–712.
16 Gu Z, Nicolae D, Lu HH, Li WH: Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet 2002;18:609–613.
17 Zhang Z, Gu J, Gu X: How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? Trends Genet 2004;20:403–407.
18 Kafri R, Bar-Even A, Pilpel Y: Transcription control reprogramming in genetic backup circuits. Nat Genet 2005;37:295–299.
19 Chang L, Khoo B, Wong L, Tropepe V: Genomic sequence and spatiotemporal expression comparison of zebrafish *mbx1* and its paralog, *mbx2*. Dev Genes Evol 2006;216:647–654.
20 Jimenez-Delgado S, Crespo M, Permanyer J, Garcia-Fernandez J, Manzanares M: Evolutionary genomics of the recently duplicated amphioxus Hairy genes. Int J Biol Sci 2006;2:66–72.
21 Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B: Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol 2004;21:1146–1151.
22 Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, et al: Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 2004;431:946–957.
23 Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y: Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci USA 2004;101:1638–1643.
24 Meyer A, Van de Peer Y: From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). BioEssays 2005;27:937–945.
25 Postlethwait JH, Yan YL, Gates MA, Horne S, Amores A, et al: Vertebrate genome evolution and the zebrafish gene map. Nat Genet 1998;18:345–349.
26 Van de Peer Y, Taylor JS, Meyer A: Are all fishes ancient polyploids? J Struct Funct Genomics 2003;3:65–73.
27 Wittbrodt J, Meyer A, Schartl M: More genes in fish? BioEssays 1998;20:511–515.
28 Van de Peer Y, Taylor JS, Braasch I, Meyer A: The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J Mol Evol 2001;53:436–446.
29 Altschmied J, Delfgaauw J, Wilde B, Duschl J, Bouneau L, et al: Subfunctionalization of duplicate *mitf* genes associated with differential degeneration of alternative exons in fish. Genetics 2002;161:259–267.
30 Bollig F, Mehringer R, Perner B, Hartung C, Schafer M, et al: Identification and comparative expression analysis of a second *wt1* gene in zebrafish. Dev Dyn 2006;235:554–561.
31 Volff JN: Genome evolution and biodiversity in teleost fish. Heredity 2005;94:280–294.
32 Winkler C, Schafer M, Duschl J, Schartl M, Volff JN: Functional divergence of two zebrafish midkine growth factors following fish-specific gene duplication. Genome Res 2003;13:1067–1081.
33 Postlethwait J, Amores A, Cresko W, Singer A, Yan YL: Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends Genet 2004;20:481–490.
34 Martinez-Barbera JP, Toresson H, Da Rocha S, Krauss S: Cloning and expression of three members of the zebrafish Bmp family: *Bmp2a*, *Bmp2b* and *Bmp4*. Gene 1997;198:53–59.
35 Imboden M, Devignot V, Goblet C: Phylogenetic relationships and chromosomal location of five distinct glycine receptor subunit genes in the teleost *Danio rerio*. Dev Genes 2001;211:415–422.

36 Ekker M, Akimenko MA, Allende ML, Smith R, Drouin G, et al: Relationships among msx gene structure and function in zebrafish and other vertebrates. Mol Biol Evol 1997;14:1008–1022.
37 Nornes S, Clarkson M, Mikkola I, Pedersen M, Bardsley A, et al: Zebrafish contains two *pax6* genes involved in eye development. Mech Dev 1998;77:185–196.
38 Force A, Shashikant C, Stadler P, Amemiya CT: Comparative genomics, *cis*-regulatory elements, and gene duplication. Methods Cell Biol 2004;77:545–561.
39 Laforest L, Brown CW, Poleo G, Geraudie J, et al: Involvement of the sonic hedgehog, *patched 1* and *bmp2* genes in patterning of the zebrafish dermal fin rays. Development 1998;125:4175–4184.
40 Joyner AL, Martin GR: *En-1* and *En-2*, two mouse genes with sequence homolog to the *Drosophila engrailed* gene: expression during embryogenesis. Genes Dev 1987;1:29–38.
41 Ensembl genome browser [http://www.ensembl.org]
42 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
43 JGI [http://genome.jgi-psf.org]
44 Notredame C, Higgins DG, Heringa J: T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302:205–217.
45 Felsenstein J: PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics 1989;5:164–166.
46 Smith TF, Waterman MS: Comparison of biosequences. Adv Appl Math 1981;2:482–489.
47 Van Hellemont R, Monsieurs P, Thijs G, De Moor B, Van de Peer Y, Marchal K: A novel approach to identifying regulatory motifs in distantly related genomes. Genome Biol 2005;6:R113.1–R113.18.
48 Wingender E, Chen X, Fricke E, Geffers R, Hehl R, et al: The TRANSFAC system on gene expression regulation. Nucleic Acids Res 2001;29:281–283.
49 Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B: Toucan: deciphering the *cis*-regulatory logic of coregulated genes. Nucleic Acids Res 2003;31:1753–1764.
50 Coessens B, Thijs G, Aerts S, Marchal K, De Smet F, et al: INCLUSive: a web portal and service registry for microarray and regulatory sequence analysis. Nucleic Acids Res 2003;31:3468–3470.
51 PubMed [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed]
52 Marchal K, Thijs G, De Keersmaecker S, Monsieurs P, De Moor B, Vanderleyden J: Genome-specific higher-order background models to improve motif detection. Trends Microbiol 2003;11:61–66.
53 Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, et al: A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 2001;17:1113–1122.
54 Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, et al: A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol 2002;9:447–464.
55 Hogan BL: Bone morphogenetic proteins: multifunctional regulators of vertebrate development. Genes Dev 1996;10:1580–1594.
56 West-Mays JA, Zhang J, Nottoli T, Hagopian-Donaldson S, Libby D, et al: AP-2alpha transcription factor is required for early morphogenesis of the lens vesicle. Dev Biol 1999;206:46–62.
57 Sivak JM, West-Mays JA, Yee A, Williams T, Fini ME: Transcription factors Pax6 and AP-2alpha interact to coordinate corneal epithelial repair by controlling expression of matrix metalloproteinase gelatinase B. Mol Cell Biol 2004;24:245–257.
58 Ingham PW, McMahon AP: Hedgehog signaling in animal development: paradigms and principles. Genes Dev 2001;15:3059–3087.
59 Shibata R, Misonou H, Campomanes CR, Anderson AE, Schrader LA, et al: A fundamental role for KChIPs in determining the molecular properties and trafficking of Kv4.2 potassium channels. J Biol Chem 2003;278:36445–36454.

Kathleen Marchal
Department of Microbial and Molecular Systems (CMPG)
K.U.Leuven, Kasteelpark Arenberg 20
B-3001 Heverlee, Leuven (Belgium)
Tel. +32169685, Fax +3216321963, E-Mail Kathleen.marchal@biw.kuleuven.be

Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 101–118

Evolution of Gene Function on the X Chromosome Versus the Autosomes

N.D. Singh, D.A. Petrov

371 Serra Mall, Stanford, Calif., USA

Abstract

Sex chromosomes have arisen from autosomes many times over the course of evolution. This process generates chromosomal heteromorphy between the sexes, which has important implications for the evolution of coding and noncoding sequences on the sex chromosomes versus the autosomes. The formation of sex chromosomes from autosomes involves a reduction in gene dosage, which can modify properties of selection pressure on sex-linked genes. This transition also generates differences in the effective population size and dominance characteristics of novel mutations on the sex chromosome versus the autosomes. All of these changes may affect both patterns of in situ gene evolution and the rates of interchromosomal gene duplication and movement. Here we present a synopsis of the current understanding of the origin of sex chromosomes, theoretical context for differences in rates and patterns of molecular evolution on the X chromosome versus the autosomes, as well as a summary of empirical molecular evolutionary data from *Drosophila* and mammalian genomes.

Copyright © 2007 S. Karger AG, Basel

Origin of Sex Chromosomes

Sex chromosomes are thought to be derived from an ancient pair of autosomes, have arisen independently several times over the course of evolution (for review see [1]), and are one of the mechanisms by which sex is determined. Sex chromosomes are generally morphologically and genetically distinct, though in some cases of recently formed sex chromosomes, the homologous neo-sex chromosomes have not had time to differentiate significantly and are thus quite similar. The heterogametic sex can be either male or female; in the case of female heterogamety the sex chromosomes are denoted Z and W while in the case of male heterogamety they are referred to as X and Y. Both types of sex chromosome evolution have arisen in diverse lineages, though the XX/XY system of

male heterogamety appears to be more common, and is found in plants, mammals, insects, amphibians, reptiles, and several groups of teleost fish, among others [2, 3]. Birds and butterflies represent the best-studied examples of the ZZ/ZW system, though this gonosomal system is also found in plants, insects, amphibians, reptiles, as well as fish [2, 3]. For the sake of simplicity, the remainder of this discussion will be in the context of male heterogametic systems.

In spite of the independent origins of these sex chromosomes in diverse taxa, there are striking similarities among them. Notably, lack of recombination in the heterogametic sex and the erosion of the Y chromosome appear to be hallmarks of sex chromosome evolution [1–4]. This degeneration of the Y typically involves gene loss as well as chromatin state transitions, and as a result, Y chromosomes are typically smaller than their homologous counterparts and are largely heterochromatic [2]. That there are common features among evolutionarily distinct sex chromosomes may be suggestive of a general framework for sex chromosome evolution.

The current model (for review see [1]) for the evolution of sex chromosomes presumes that the sex chromosomes originate from a homologous pair of freely recombining autosomes. The first step in the transition between the autosomal and sex-linked state is the restriction of recombination between the proto-X and the proto-Y, though explanations for why such suppression evolves remain elusive [2]. Most likely, if alleles contributing to sex-determination arise, recombination between these genes will be deleterious, as it would generate maladapted sexual phenotypes [4]. Following this crucial step, natural selection may continue to favor tight linkage of the genes involved in sex determination. Selection may also promote the linkage of these sex-determining loci with sexually antagonistic genes, which are advantageous in one sex while deleterious in the other, as genes with sexually antagonistic consequences that cosegregate with the appropriate sex-determining locus will enjoy a selective advantage over those that segregate independently. This process will differentiate the sex chromosomes further, which will in turn generate selective pressure for the suppression of recombination over an increasingly large window [4]; inversions and rearrangements, which have been implicated in the evolution of the human sex chromosomes [5, 6], may play a role at this stage.

Once the X and the Y are differentiated and recombination between them has reached sufficiently low levels, the Y chromosome begins to erode genetically. There are five main theories for this degeneration, and it is not yet clear which process is primarily governing this genetic erosion of the Y chromosome (for review see [7, 8]). One possibility is that the Y chromosome degenerates due to the effects of Muller's Ratchet [9, 10], or the continued and stochastic loss of the chromosome class containing the fewest deleterious mutations in non-recombining, finite populations. A background selection model [11] similarly

predicts a gradual decline in fitness of the Y chromosome; background selection reduces the effective population size of the non-recombining Y chromosome by removing deleterious alleles and neutral variants on this allelic background from the population, which facilitates the fixation of mildly deleterious mutants due to lowered effective population size.

Another hypothesis is 'weak-selection Hill-Robertson effect' [12], which posits that fixation probabilities of alleles are altered by fixation probabilities of nearby, linked mutations; this general model can contribute to the degeneration of the Y chromosome but the timescales required under this model may be too great to be of biological significance [8]. Genetic hitchhiking [13] may also play a role in Y chromosome degeneration [14], because fixation of novel adaptive alleles also results in fixation of all linked deleterious mutations. As the Y chromosome is unable to recombine, repeated bouts of positive selection will lead to an accumulation of deleterious Y-linked alleles. However, this model as well requires timescales that are too great to be appropriate for sex chromosome evolution. Orr's and Kim's "ruby in the rubbish" model [15] is also based on the inability of the non-recombining Y chromosome to evolve adaptively. In this model, beneficial Y-linked mutations rarely reach fixation, creating a large fitness difference between the X and Y chromosomes, which generates selective pressure to decrease the expression of these deleterious Y-linked genes. Thus, there are several non-mutually exclusive hypotheses for the processes underlying the evolution of the Y chromosome, and further work is needed to elucidate the relative roles of these forces in Y chromosome evolution.

In this review, we will focus on the implications of the degeneration of the Y chromosome for the evolution of coding sequences on the X chromosome, particularly with respect to differences in molecular evolution between the X and the autosomes. We will present an introduction to the theoretical and conceptual frameworks for understanding the evolution of X-linked sequences, as well as summaries of available experimental evidence to date.

Theoretical Predictions for Molecular Evolution of X-linked and Autosomal Genes

Three major differences between the X chromosome and the autosomes stem from the evolution of sex chromosomes from autosomes, each of which has important implications for the evolution of coding and noncoding sequences on these chromosome sets. First, as a consequence of the presence of only a single copy of the X chromosome in males, the X and the autosomes may differ with respect to effective population size. Assuming equal numbers of breeding males and breeding females, the effective population size of the

X chromosome should be 3/4 that of the autosomes, since there are only three X chromosomes for every four autosomes segregating in the population. However, other factors will serve to modulate the effective population size of the X chromosome relative to that of the autosomes such as sex-specific life history traits or differences in breeding success between males and females. In fact, if the effective number of breeding females far exceeds the effective number of breeding males, the effective population size of the X chromosome can equal or even exceed that of the autosomes [16, 17]. Effective population size is an important molecular evolutionary parameter, as increases in effective population size can increase the efficacy of natural selection on weakly adaptive or mildly deleterious mutations.

The effective haploidy of the X chromosome in males also affects the visibility of novel mutations to natural selection. If new mutations are at least partially recessive, the selective effects of novel autosomal variants can be masked by the homologous allele in heterozygous individuals. In contrast, X-linked alleles are immediately visible to selection in males, as males are hemizygous for the X chromosome. Although the X chromosome spends only 1/3 of its evolutionary history in males, the increased exposure of X-linked alleles to selection in these hemizygous individuals enhances the efficacy of both positive and negative selection [18], which can categorically alter rates of molecular evolution between the X and the autosomes.

Finally, that the sex chromosomes are formed from a pair of ancestral autosomes generates a dosage problem for X-linked genes. Indeed, the degradation of the Y chromosome involves gene loss [4, 8], and as a result, X-linked genes generally lack a functional counterpart on the Y and are thus present at one half the copy number that they were in the ancestral, autosomal state. Such a reduction in gene dosage is likely to be deleterious for many genes and a number of dosage compensation mechanisms have evolved in several lineages to remediate the effects of this dosage problem. *Drosophila* and mammals provide the most well-characterized mechanisms of dosage compensation (for review see [19, 20]); in *Drosophila*, dosage compensation is mediated by transcriptional upregulation of X-linked genes in males, whereas in mammals one copy of the X chromosome is transcriptionally downregulated in females.

Though much dosage compensation is achieved through transcriptional regulation, the reduction in gene dosage of X-linked genes may also alter the strength of selection on other mutations that could also serve to partially remediate the dosage problem. For instance, duplication or transposition events from the sex chromosome to the autosomes in some cases may be selectively favored to mitigate the reduction in gene dosage [21–23]. In addition, selection may favor increased codon bias for X-linked genes, which may increase levels of active protein [24] and thus partially compensate for the dosage problem. Even

if dosage equilibration is reached between the sexes, selective pressures may still affect the X chromosome and the autosomes differently.

The chromosomal heteromorphy which results from the evolution of sex chromosomes from autosomes can thus modify properties of selection pressure, effective population size and dominance characteristics of mutations in X-linked genes. These parameters will interact to affect rates and patterns of evolution on the X chromosome relative to the autosomes, and the relative rates of evolution of X-linked versus autosomal genes under various parameter combinations can be examined. In particular, the relative rates of evolution between the chromosome sets will depend on two properties of new mutations: dominance characteristics and sojourn times.

The coefficient of dominance of a new mutation is a key determinant of the relative rate of spread of that new allele on the X chromosome versus the autosomes. Assuming equal numbers of breeding males and breeding females, rates of evolution on the X chromosome will exceed those on the autosomes if new mutations are on average at least partially recessive [18, 25], for both small and large coefficients of selection [26]. In contrast, rates of substitution of mildly deleterious alleles on the autosomes will exceed those on the X [18]. For codominant mutations, rates of evolution should be comparable between the X and the autosomes and rates of fixation of at least partially dominant mutations on the autosomes will exceed those of the X chromosome, for mutations of both small and large selective effects [18, 26].

Sojourn time of novel mutants differs between the X and the autosomes, and this may also affect rates of evolution for non-neutral mutations between the X and the autosomes. In general, the sojourn time for new beneficial mutations will be shorter if the mutations are X-linked rather than autosomal [25, 27], though the magnitude of the difference in rates of evolution between the X and the autosomes does depend on both the relative numbers of breeding males and females as well as the coefficient of selection [25]. This inequality in transit time holds across all dominance coefficients [25, 26], and results from the greater variance in fitness in the haploid versus the diploid state. Given that changes in allele frequency are a function of variance in fitness [28], that the X chromosome spends 1/3 of its evolutionary history in the haploid state suggests that the change in frequency of a novel beneficial allele will be greater if it is X-linked.

It should be noted that all of the above theoretical predictions are predicated on selection acting on novel mutational variants. If natural selection instead predominantly operates on standing variation, with formerly deleterious mutations becoming advantageous, these predictions no longer hold. Instead, rates of adaptive evolution are slower for X-linked alleles than for autosomal alleles [18], regardless of the dominance coefficient of these alleles [29].

Theoretical predictions are also sensitive to the strength and direction of selection acting on mutant alleles in the two sexes. If selection is not presupposed to be acting equally in the two sexes and rather, if there are opposing selective pressures in males versus females, as is the case for sexually antagonistic genes, then X-linked mutations that benefit males can spread through a population under a less restrictive range of parameter values than is required for autosomal invasion [30]. The opposite appears to be true when the mutations favor females at the cost of males, at least for partially recessive mutations [18].

Patterns of In Situ Evolution of X-linked and Autosomal Genes

The above theoretical considerations suggest that under certain conditions, X-linked loci will have higher rates of adaptive evolution than autosomal loci. For this to be the case, selection must be operating on novel allelic variants, and the selective effects of these mutations must be on average at least partially recessive and be equal in the two sexes [18]. In addition, fixation of alleles under positive selection should predominate over fixation of slightly deleterious alleles. Empirical evidence in support of this 'faster-X' hypothesis is thus taken as confirmation that such assumptions are biologically reasonable.

Early testing for faster-X evolution yielded largely contradictory results. A recent comparison of rates of protein evolution in 254 coding sequences from *D. melanogaster* and *D. simulans* showed comparable rates of evolution for X-linked and autosomal genes [31]. In contrast, comparative genomic data from pairs of orthologous genes from these species as well as *D. pseudoobscura* and *D. miranda* suggest that rates of evolution on the X chromosome exceed those on the autosomes [32], and that rates of adaptive evolution between pairs of X-linked gene duplicates appear to be higher than those for duplicate gene pairs residing on the autosomes [33]. The discrepancy between these two major findings may be a consequence of experimental design; faster-X evolution predicts an increased rate of adaptive evolution for an X-linked gene relative to the rate of evolution that a gene would experience if it were autosomal. As a result, paired comparisons, either between orthologous sequences from different species or paralogous sequences within species, may be more appropriate for testing the faster-X model.

With several *Drosophila* genomes fully sequenced, investigating faster-X in this system is now possible at a large scale. Using the full genomes of *D. melanogaster*, *D. pseudoobscura*, *D. yakuba* as well as large-scale sequence data from *D. miranda*, Thornton and colleagues revisited this question at the genomic scale [34]. Using either only whole genome data from the three Drosophilids or a smaller dataset of 202 coding sequences from all four species

did not qualitatively or quantitatively affect the results; rates of protein evolution are not significantly different between the X and the autosomes. The authors suggest that this lack of support for faster-X evolution indicates that either new mutations are not on average partially recessive or that adaptive evolution originates from mutation-selection equilibrium [34].

Results from mammalian genomes are more consistent, with several studies offering evidence in support of a faster-X model of evolution. A whole genome comparison of the human and chimpanzee genomes reveals rates of protein evolution (estimated as Ka/Ks, or the ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site) of X-linked genes exceeding rates of evolution on the autosomes [35]. In addition, a scan for positively selected genes in these genomes revealed an enrichment of X-linked genes, suggesting that X-linked genes have an increased tendency for adaptive evolution relative to autosomal genes, which is also consistent with the faster-X model [36]. Similarly, inferring selective events using linkage disequilibrium among single nucleotide polymorphisms results in a 2-fold enrichment of putatively selected loci on the X chromosome in humans [37]. These genomic phenomena are also recapitulated within smaller functional categories of genes; sex-linked mammalian sperm proteins, for instance, evolve more rapidly than autosomal sperm proteins [38, 39], and X-linked testis-expressed genes have higher rates of evolution (normalized to account for local mutation rate) than those on the autosomes [40], as do X-linked testis-expressed homeobox genes [41]. While these data do suggest that X-linked genes may indeed evolve more rapidly than autosomal genes in mammalian genomes, there is a possibility that ascertainment bias also plays a role in generating these patterns; an overrepresentation of X-linked genes in rapidly evolving proteins, for instance, could reflect differences in gene complements between the X and the autosomes rather than higher rates of adaptive evolution for X-linked genes.

In addition to examining patterns of interspecific divergence of X-linked versus autosomal genes, patterns of variability within species at sex-linked and autosomal loci can also shed light on the forces contributing to the evolution of coding and noncoding sequences on these chromosome sets. Adaptive evolution and purifying selection will both contribute to levels of intraspecific variation via the effects of genetic hitchhiking and background selection, respectively; differences in the relative contributions of these evolutionary processes in X-linked versus autosomal genes may manifest as differences in standing levels of variation. Importantly, background selection and hitchhiking models make distinct predictions regarding the relative levels of diversity of X-linked and autosomal genes. Background selection models predict higher levels of neutral variation on the X chromosome [11, 27, 42, 43]. The reduction

of variation at neutral sites due to the removal of deleterious alleles and linked neutral variants by purifying selection is most pronounced when deleterious alleles reach high frequencies. Given that purifying selection is more effective on the X chromosome due to the hemizygosity of the X chromosome in males, deleterious mutants are maintained at lower population frequencies if they are X-linked than they would be if they were autosomal. In essence, the X chromosome has a larger effective number of deleterious-mutation free chromosomes, and as a consequence, increased levels of standing neutral variation, relative to the autosomes.

In contrast to the background selection model, genetic hitchhiking may lead to lower levels of polymorphism on the X chromosome versus the autosomes. Because the sojourn time of new adaptive mutations is shorter for X-linked versus autosomal genes [25, 27], there may be fewer recombinational opportunities during a selective sweep of an X-linked gene than there would be in an autosomal gene. In addition, if new beneficial mutations are at least partially recessive on average, X-linked genes simply evolve more rapidly from adaptive evolution, which would result in an increase in the number of selective sweeps of X-linked alleles over autosomal alleles per unit time. Theoretical results suggest that such a hitchhiking model will yield lower diversity on the X chromosome than on the autosomes if new beneficial mutations are partially recessive in systems such as *Drosophila* in which there is no recombination in males, and under a broader range of dominance coefficients if there is recombination in males, as is the case in humans [26].

As these models make different predictions with respect to expected levels of neutral sequence variation on the X and the autosomes, comparing X-linked and autosomal polymorphism can provide insight into the relative roles of background selection versus hitchhiking models. There are numerous studies of levels of molecular polymorphism in *D. melanogaster* and *D. simulans* (for review see [44]). Within *D. simulans*, levels of diversity are consistently lower at X-linked loci [42, 45–47]. In *D. melanogaster*, differences in sequence variation between the X and the autosomes seem heavily dependent on population. For ancestral African populations, X-linked diversity levels are consistently higher than expected under the assumption of equal numbers of breeding males and breeding females, while X-linked diversity appears to be depressed in derived populations of this species [45, 48, 49].

Studies of variation in other taxa have revealed similar patterns. In humans, the densities of single nucleotide polymorphisms and microsatellite markers are considerably lower on the X chromosome relative to the autosomes [50, 51], and noncoding sequence diversity as well as microsatellite variability also seem to be reduced on the human X chromosome [51–53]. The density of SSLP markers as well as polymorphism at these loci in mouse are also depressed for X-linked

loci relative to autosomal loci [54]; this X-specific deficit in polymorphic markers is also found in rat [55]. In chicken as well as two flycatcher species, polymorphism data from intronic sequences also support a reduction in diversity of Z-linked alleles relative to autosomal levels [56, 57].

On balance, it appears as though positive selection does play a role in the evolution of the sex chromosomes. While specifically testing for faster-X in *Drosophila* has generated inconsistent results, patterns of X chromosome evolution in mammals appear wholly consistent with the faster-X model. The discrepancies within the *Drosophila* studies may in fact suggest that positive selection is comparatively rare in this system (although see [58]), or may result from the breakdown of one or more of the assumptions in the faster-X model. Intraspecific patterns of sequence variability are more consistent overall among taxa, with a general trend towards reduced polymorphism of the X (or Z) chromosomes. Such a reduction is consistent with a model of genetic hitchhiking or reduced effective population size, and may thus implicate positive selection in the evolution of the X chromosome.

It should be noted that purifying selection, or selection against deleterious alleles, may be more efficient on the X chromosome as well, in accordance with theoretical predictions [18]. Evidence in support of this model stems largely from studies of the evolution of codon bias on the X chromosome and the autosomes. Codon bias refers to the unequal usage of synonymous codons in protein coding sequences, and is thought to be maintained by the balance among mutation, random genetic drift, and selection on translational efficiency/accuracy [59–62].

Codon bias of X-linked genes appears to be higher than codon bias of genes on the autosomes in *Drosophila* [63–65] and *C. elegans* [65]. This increase in codon bias on the X chromosome in these two systems is not mediated by other known correlates of codon bias such as recombination rate, protein length, or level of gene expression. In addition, the X-specific elevation in codon bias does not result from the identities or functions of the genes residing on this chromosome, as comparisons of codon bias in pairs of X-linked and autosomal duplicate genes in *D. melanogaster* and *C. elegans*, as well as pairs of orthologous genes involved in an X-autosome translocation in *D. melanogaster* and *D. pseudoobscura* also support the increase in codon bias on the X chromosome [65]. Thus, the increase in codon bias on the X chromosome in *Drosophila* and *C. elegans* appears to be due entirely to X-linkage, and is thus consistent with an increased efficacy of purifying selection on the X chromosome.

Importantly, while many of the above observations are indeed consistent with the increased efficacy of both directional and purifying selection on the X chromosome, which is suggestive of an in situ evolutionary model, it is also possible that the observed differences between rates and patterns of molecular

evolution are due to other forces shaping the properties of the resident genes on these different chromosomes. External forces may be of particular importance in light of observed differences between the X and the autosomes with respect to gene content and gene movement, which are the focus of the remainder of this discussion.

Gene Complements of the X Chromosome Versus the Autosomes

In addition to predicting differences in rates of molecular evolution between the X and the autosomes, theoretical models further suggest that differences in gene content may evolve between the X chromosome and the autosomes. In particular, genes with different selective effects in males versus females may accumulate differently on the X versus the autosomes, thus shaping the complements of the genes residing on these chromosome sets. More precisely, genes that benefit one sex at the cost of the other, known as sexually antagonistic genes, can accumulate faster on the X chromosome than on the autosomes under certain conditions [30]. For both partially dominant and partially recessive alleles, sexually antagonistic alleles that are beneficial in males though detrimental to females enjoy higher fixation rates if they are X-linked, though the requisite conditions differ somewhat between these classes of dominance coefficients [30]. Mutations in X-linked genes that benefit females at the cost of males can also increase in frequency more rapidly than comparable autosomal mutations, particularly if these mutations are dominant [30]. Thus, rates of accumulation of sexually antagonistic alleles benefiting females or males on the X chromosome can exceed those rates on the autosomes under a variety of conditions, which may play a role in shaping the gene content of the X versus the autosomes.

In *Drosophila*, there appears to be a relative dearth of male-biased genes and an enrichment of female-biased genes on the X chromosome. The genomic distribution of secreted accessory gland proteins, for example, which are heavily implicated in male reproduction, is shifted significantly away from the X chromosome [21]. In addition, genes with male-biased patterns of germline and somatic gene expression are comparatively rare on the *Drosophila* X [66, 67], and genes with female-biased germline expression are enriched on the X relative to expectation [67].

Similar patterns have been documented in *C. elegans*. In addition to genes with germline expression, male-biased germline expressed genes are less likely to be found on the X chromosome [68, 69] though genes with hermaphrodite somatic-biased expression are enriched on the X [69]. Although seemingly unrelated to the sexual antagonism model, the gene complements of the X and

autosomes in *C. elegans* also differ in other respects; the X chromosome is also relatively devoid of genes essential for basic cellular and developmental processes in the embryo [70].

The distribution of genes with sexually antagonistic consequences in mammalian systems is less straightforward. While genes with male-specific expression patterns consistently appear to be less frequent on the X chromosome in humans, there does not appear to be an overrepresentation of X-linked female-specific genes [71]. Genes with sex- and reproduction-related functions appear to be enriched on the human X as well [72]. In mouse, both a dearth of male-specific and an overabundance of female-specific genes on the X have been documented [73], though genes expressed in spermatogonia in this system are more likely to be X-linked than autosomal [74].

One component of the explanation for the conflicting results from *Drosophila*, *C. elegans*, and mammals in relation to one another as well as to theoretical predictions [30] is related to the inactivation of the X chromosome, which occurs during meiosis in the male germline of mammals and many insect taxa [75]. Thus, genes required in late spermatogenesis are at a selective disadvantage if they are X-linked rather than autosomal, although the same is not the case for genes expressed in the male germline prior to inactivation. Thus, we might not expect to find an accumulation of X-linked male-biased genes for those genes that are expressed after X inactivation (for review see [76]). This is supported by data from mouse, which indicate that the X has a depletion of male-biased genes that are expressed late in spermatogenesis and an enrichment of male-biased genes expressed earlier [73], thus supporting the intersection of the sexual-antagonism [30] and X-inactivation models. Moreover, of the 26 genes identified as acting in late spermatogenesis none appears to be X-linked [77].

Patterns of Gene Traffic on the X and the Autosomes

In addition to differing with respect to rates and patterns of molecular evolution as well as gene content, the X and the autosomes also differ with respect to patterns of gene movement via retrotransposition. While duplicate genes can arise from several mutational mechanisms such as small-scale duplication or whole genome duplication, novel genes created through retrotransposition are somewhat more readily identifiable. The process leading to the formation of these retroposed gene duplicates involves reverse transcription of the mRNA from the parental gene and subsequent insertion of this new DNA sequence into an ectopic location in the genome. As a direct consequence, duplicate genes generated through this mechanism bear signatures of this process, which include the lack of introns, poly-A tracts as well as direct flanking repeats.

These latter two characteristics may erode over evolutionary time due to the accumulation of single nucleotide, insertion and deletion mutations. Although there do appear to be cases in which retroposed genes have recruited new introns [78], the lack of introns is a more stable feature of retrotransposed genes, and is thus generally used as a criterion for identifying retroposed genes.

Studies of the fate of retrotransposed genes have been carried out extensively in *Drosophila* and mammals (for review see [79, 80]). In *Drosophila*, there appears to be a significant excess of retrotransposition of X-linked genes to autosomal locations [77]. This excess of X-linked retrogene origination is similarly documented in mammals [78, 81], although in this system the X chromosome also disproportionately recruits duplicate genes arising through retrotransposition [81]. While the X chromosome also shows increased recruitment of retropseudogenes relative to expectation, which implicates a mutational bias, this mutational explanation is not wholly sufficient to explain the patterns of gene traffic of functional retrogenes in mammals [81].

Interestingly, a large fraction of these retrogenes with X-linked parents derive testis-specific expression patterns [77, 81]. In *Drosophila*, five out of six retrogenes originating from the X chromosome are expressed in testis while their parental gene is not (see [77]). In mammals, a higher percentage of X-originating autosomal retrogenes are expressed in testis than autosomal retrogenes that originate from autosomal genes [81], though, it is not yet clear whether these testis-biased expression patterns are predominantly derived or ancestral. More recent analysis of functional retrogenes in the human genome revealed seven functional duplicate genes arising through retrotransposition, three of which originated from X-linked genes, and all of which had acquired novel testis-biased or testis-specific expression patterns [82]. Larger sampling of retrogenes in humans also supports the hypothesis that retrogenes tend to be expressed in testis [78].

Beyond gene retrotransposition, rates of interchromosomal gene movement for all mechanisms of genic translocation also vary substantially between the X chromosome and the autosomes in *Drosophila* [Davis, Singh and Petrov, unpublished]. By comparing physical map locations between pairs of orthologous genes in *D. pseudoobscura* and *D. melanogaster*, asymmetries in gene movement rates between the sex chromosomes and the autosomes can be explored, with emphasis on the newly formed sex chromosomes in the *D. pseudoobscura* lineage. Preliminary analysis of comparative map locations of orthologous genes in these two species has provided tantalizing evidence that rates of gene movement differ between the sex chromosomes and the autosomes [Davis, Singh and Petrov, unpublished]. In particular, the autosome-X translocation has led to a strong bias toward overall gene loss from the neo-X chromosome in *D. pseudoobscura*. Specifically, it was estimated using a maximum

likelihood framework that while the rate of gene emigration from the neo-X increased by as much as 8-fold, the rate of gene immigration to the neo-X declined to undetectable levels. In addition, these preliminary results suggest that rates of gene movement between the ancestral X and autosomes are higher than the rates of interautosomal gene movement.

There are several models that have been put forward to explain patterns of gene movement between the X and the autosomes, none of which can fully account for all of the described patterns (for review see [79]). The X-inactivation hypothesis suggests that selection favors autosomal locations for retroposed genes with functions requiring expression during male meiosis because of the inactivation of the X chromosome in male germline cells. A related hypothesis, the SAXI hypothesis [76], predicts the redistribution of genes functioning in late spermatogenesis from the X chromosome to the autosomes and the gradual demasculinization of the X chromosome perhaps as a consequence of interactions among sexually antagonistic alleles. Finally, formal population genetic models suggest that mutations that are at least partially dominant may accumulate at higher rates on the autosomes [18], and it seems likely that mutations in genes with sexually antagonistic or sex-limited effects would be predominantly gain-of-function mutations and therefore be at least partially dominant [67]. Each of these models can explain certain aspects of the observed data, but no single model appears to be consistent with all previous reports. Consequently, further investigation of the relative importance of these models in generating the observed patterns of gene movement among chromosomes is warranted.

Summary

Sex chromosomes have evolved from autosomes independently in diverse lineages. The prominent feature of the transition between the ancestral, autosomal state and the sex chromosome state is the reduction in gene dosage of sex-linked alleles in the heterogametic sex. The effective haploidy of the sex chromosome in heterogametes is thus the foundation for major differences between the sex chromosomes and the autosomes, which are of tremendous consequence for the evolution of these sets of chromosomes. A great deal of theoretical attention has been devoted to the evolution of sex-linked versus autosomal alleles [18, 25]. These results suggest that rates of evolution may differ between the X and the autosomes both with respect to adaptive evolution and purifying selection, although the magnitude and direction of the difference depends on parameters such as the coefficient of dominance [18, 25] and on other features of the model such as, for example, whether evolution primarily acts on novel

mutational variants or standing variation [29]. In addition, theoretical considerations of genes with different selective effects in the two sexes predict that gene complements between the X and the autosomes may differ depending on the fitness consequences in each sex and the dominance of mutations affecting these genes [30].

The bulk of the data on contrasting patterns of X chromosomal and autosomal evolution comes from *Drosophila* and mammals. While the data are somewhat conflicting, overall there appears to be some evidence in support of a faster-X model of evolution, suggesting that positive selection plays a role in the evolution of sex chromosomes, although may not be a prominent feature of X-linked genes overall. Increased rates of adaptive evolution for X-linked genes may indeed be restricted to genes that are evolving rapidly under positive selection. Within species polymorphism data from X-linked and autosomal loci are also consistent with increased action of positive selection on the X as well [44, 50–53, 56, 57]. Data from codon bias evolution studies also suggest that purifying selection is more effective on the X chromosome than on the autosomes [63–65]. Together, these data lend support to the recessivity of both beneficial and deleterious alleles.

Although the gene complements of the X and the autosomes tend to differ, there do not appear to be any systematic trends across taxa. In *Drosophila* and *C. elegans*, the X chromosome is relatively devoid of male-biased genes while female or hermaphrodite-biased genes appear to be overrepresented [21, 66–69]. The mammalian X also shows a deficiency of genes expressed late in spermatogenesis, and shows a putative enrichment of female-biased genes as well as an overrepresentation of male-biased genes expressed early in the germline [73].

Similarities and differences between mammals and *Drosophila* are also found with respect to patterns of gene traffic of X-linked and autosomal genes, as well as in the functional characteristics of these retrogenes (for review see [79]). While in both mammals and *Drosophila* the X chromosome disproportionately exports new retrogenes, in mammals the X chromosome also recruits retrogenes in excess of expectation. Retrogenes in both systems tend to be expressed in testis, although it remains to be seen how much of this effect is due to acquisition of novel expression pattern in the derived retroposed gene. While patterns of gene traffic in general appear to differ between the X and the autosomes in *Drosophila* [Davis, Singh and Petrov, unpublished], it remains to be seen whether this is also the case in mammalian genomes. Although several models have been proposed to explain these patterns, none can sufficiently account for all of the observations.

Thus, the sex chromosomes of diverse species share several salient evolutionary features. The similarities among patterns of X-linked and autosomal evolution in these systems likely result from similar evolutionary forces acting

on sex-linked genes in spite of their independent origin. However, there are marked differences between the systems presented here, which may speak to differences in the relative roles of the evolutionary forces of mutation, random genetic drift, and natural selection among these organisms. Clearly, the appropriate framework for understanding the evolution of coding and noncoding sequences on sex chromosomes and the autosomes will integrate general features of sex chromosome evolution as well as lineage-specific effects.

References

1 Charlesworth B: The evolution of sex chromosomes. Science 1991;251:1030–1033.
2 Bull JJ: Evolution of Sex Determining Mechanisms. Benjamin/Cummings Pub. Co. Advanced Book Program, Menlo Park 1983.
3 Solari AJ: Sex Chromosomes and Sex Determination in Vertebrates. CRC Press, Boca Raton 1994.
4 Charlesworth B: The evolution of chromosomal sex determination and dosage compensation. Curr Biol 1996;6:149–162.
5 Marais G, Galtier N: Sex chromosomes: how X-Y recombination stops. Curr Biol 2003;13: R641–R643.
6 Lahn BT, Page DC: Four evolutionary strata on the human X chromosome. Science 1999;286: 964–967.
7 Charlesworth B: Model for evolution of Y chromosomes and dosage compensation. Proc Natl Acad Sci USA 1978;75:5618–5622.
8 Charlesworth B, Charlesworth D: The degeneration of Y chromosomes. Phil Trans R Soc Lond B Biol Sci 2000;355:1563–1572.
9 Muller HJ: The relation of recombination to mutational advance. Mutat Res 1964;1:2–9.
10 Felsenstein J: The evolutionary advantage of recombination. Genetics 1974;78:737–756.
11 Charlesworth B, Morgan MT, Charlesworth D: The effect of deleterious mutations on neutral molecular variation. Genetics 1993;134:1289–1303.
12 McVean GAT, Charlesworth B: The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics 2000;155:929–944.
13 Maynard Smith J, Haigh J: The hitch-hiking effect of a favourable gene. Genet Res 1974;23: 23–35.
14 Rice WR: Genetic hitch-hiking and the evolution of reduced genetic activity of the Y chromosome. Genetics 1987;116:161–167.
15 Orr HA, Kim Y: An adaptive hypothesis for the evolution of the Y chromosome. Genetics 1998;150:1693–1698.
16 Caballero A: On the effective size of populations with separate sexes, with particular reference to sex-linked genes. Genetics 1995;139:1007–1011.
17 Laporte V, Charlesworth B: Effective population size and population subdivision in demographically structured populations. Genetics 2002;162:501–519.
18 Charlesworth B, Coyne JA, Barton NH: The relative rates of evolution of sex chromosomes and autosomes. Am Nat 1987;130:113–146.
19 Baker BS, Gorman M, Marin I: Dosage compensation in *Drosophila*. Ann Rev Genet 1994;28: 491–521.
20 Marin I, Siegal ML, Baker BS: The evolution of dosage-compensation mechanisms. Bioessays 2000;22:1106–1114.
21 Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF: Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc Natl Acad Sci USA 2001;98:7375–7379.

22 Bachtrog D, Charlesworth B: On the genomic location of the *exuperantia1* gene in *Drosophila miranda* – the limits of in situ hybridization experiments. Genetics 2003;164:1237–1240.
23 Yi S, Charlesworth B: A selective sweep associated with a recent gene transposition in *Drosophila miranda*. Genetics 2000;156:1753–1763.
24 Carlini DB, Stephan W: In vivo introduction of unpreferred synonymous codons into the *Drosophila Adh* gene results in reduced levels of ADH protein. Genetics 2003;163:239–243.
25 Avery PJ: The population genetics of haplo-diploids and X-linked genes. Genet Res 1984;44: 321–341.
26 Betancourt AJ, Kim Y, Orr HA: A pseudohitchhiking model of X vs. autosomal diversity. Genetics 2004;168:2261–2269.
27 Aquadro CF, Begun DJ, Kindahl EC: Selection, recombination, and DNA polymorphism in *Drosophila*; in Golding B (ed): Non-Neutral Evolution. Chapman & Hall, London, 1994, pp 46–56.
28 Fisher RA: The Genetical Theory of Natural Selection. Dover, New York, 1958.
29 Orr HA, Bentancourt AJ: Haldane's sieve and adaptation from the standing genetic variation. Genetics 2001;157:875–884.
30 Rice WR: Sex chromosomes and the evolution of sex dimorphism. Evolution 1984;38:735–742.
31 Betancourt AJ, Presgraves DC, Swanson WJ: A test for faster X evolution in *Drosophila*. Mol Biol Evol 2002;19:1816–1819.
32 Counterman BA, Ortiz-Barrientos D, Noor MA: Using comparative genomic data to test for fast-X evolution. Evolution 2004;58:656–660.
33 Thornton K, Long M: Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. Mol Biol Evol 2002;19:918–925.
34 Thornton K, Bachtrog D, Andolfatto P: X chromosomes and autosomes evolve at similar rates in *Drosophila*: no evidence for faster-X protein evolution. Genome Res 2006;16:498–504.
35 Lu J, Wu CI: Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc Natl Acad Sci USA 2005;102:4063–4067.
36 Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, et al: A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 2005;3:976–985.
37 Wang ET, Kodama G, Baldi P, Moyzis RK: Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proc Natl Acad Sci USA 2006;103:135–140.
38 Torgerson DG, Singh RS: Sex-linked mammalian sperm proteins evolve faster than autosomal ones. Mol Biol Evol 2003;20:1705–1709.
39 Torgerson DG, Singh RS: Enhanced adaptive evolution of sperm-expressed genes on the mammalian X chromosome. Heredity 2006;96:39–44.
40 Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, et al: Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 2005;309:1850–1854.
41 Wang XX, Zhang JZ: Rapid evolution of mammalian X-linked testis-expressed homeobox genes. Genetics 2004;167:879–888.
42 Begun DJ, Whitley P: Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc Natl Acad Sci USA 2000;97:5960–5965.
43 Charlesworth B: Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet Res 1996;68:131–149.
44 Mousset S, Derome N: Molecular polymorphism in *Drosophila melanogaster* and *D. simulans*: what have we learned from recent studies? Genetica 2004;120:79–86.
45 Andolfatto P: Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol Biol Evol 2001;18:279–290.
46 Schoefl G, Schloetterer C: Patterns of microsatellite variability among X chromosomes and autosomes indicate a high frequency of beneficial mutations in non-african *D. simulans*. Mol Biol Evol 2004;21:1384–1390.
47 Irvin SD, Wetterstrand KA, Hutter CM, Aquadro CF: Genetic variation and differentiation at microsatellite loci in *Drosophila simulans*: evidence for founder effects in new world populations. Genetics 1998;150:777–790.
48 Sheldahl LA, Weinreich DM, Rand DM: Recombination, dominance and selection on amino acid polymorphism in the *Drosophila* genome: contrasting patterns on the X and fourth chromosomes. Genetics 2003;165:1195–1208.

49 Kauer M, Zangerl B, Dieringer D, Schlotterer C: Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. Genetics 2002;160:247–256.
50 Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001;409: 928–933.
51 Dib C, Faure S, Fizames C, Samson D, Drouot N, et al: A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 1996;380:152–154.
52 Payseur BA, Cutter AD, Nachman MW: Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol Biol Evol 2002;19:1143–1153.
53 Payseur BA, Nachman MW: Natural selection at linked sites in humans. Gene 2002;300:31–42.
54 Dietrich WF, Miller J, Steen R, Merchant MA, Damron-Boles D, et al: A comprehensive genetic map of the mouse genome. Nature 1996;380:149–152.
55 Jacob HJ, Brown DM, Bunker RK, Daly MJ, Dzau VJ, et al: A genetic linkage map of the laboratory rat, *Rattus norvegicus*. Nat Genet 1995;9:63–69.
56 Borge T, Webster MT, Andersson G, Saetre GP: Contrasting patterns of polymorphism and divergence on the Z chromosome and autosomes in two *Ficedula* flycatcher species. Genetics 2005;171: 1861–1873.
57 Sundstrom H, Webster MT, Ellegren H: Reduced variation on the chicken Z chromosome. Genetics 2004;167:377–385.
58 Andolfatto P: Adaptive evolution of non-coding DNA in *Drosophila*. Nature 2005;437: 1149–1152.
59 Sharp PM, Li WH: An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 1986;24:28–38.
60 Bulmer M: The selection-mutation-drift theory of synonymous codon usage. Genetics 1991;129: 897–908.
61 Akashi H: Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. Gene 1997;205:269–278.
62 McVean GAT, Charlesworth B: A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet Res 1999;74:145–158.
63 Hambuch TM, Parsch J: Patterns of synonymous codon usage in *Drosophila melanogaster* genes with sex-biased expression. Genetics 2005;170:1691–1700.
64 Comeron JM, Kreitman M, Aguade M: Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics 1999;151:239–249.
65 Singh ND, Davis JC, Petrov DA: X-linked genes evolve higher codon bias in *Drosophila* and *Caenorhabditis*. Genetics 2005;171:145–155.
66 Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, et al: Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. Science 2003;299:697–700.
67 Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL: Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science 2003;300:1742–1745.
68 Reinke V, Smith HE, Nance J, Wang J, Van Doren C, et al: A global profile of germline gene expression in *C. elegans*. Mol Cell 2000;6:605–616.
69 Reinke V, Gil IS, Ward S, Kazmer K: Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. Development 2004;131:311–323.
70 Piano F, Schetter AJ, Morton DG, Gunsalus KC, Reinke V, et al: Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C. elegans*. Curr Biol 2002;12:1959–1964.
71 Lercher MJ, Urrutia AO, Hurst LD: Evidence that the human X chromosome is enriched for male-specific but not female-specific genes. Mol Biol Evol 2003;20:1113–1116.
72 Saifi GM, Chandra HS: An apparent excess of sex and reproduction-related genes on the human X chromosome. Proc R Soc Biol Sci B 1999;270:53–59.
73 Khil PP, Smirnova NA, Romanienko PJ, Camerini-Otero RD: The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. Nat Genet 2004;36:642–646.
74 Wang PJ, McCarrey JR, Yang F, Page DC: An abundance of X-linked genes expressed in spermatogonia. Nat Genet 2001;27:422–426.

75 Lifschytz E, Lindsley DL: The role of X chromosome inactivation during spermatogenesis (Drosophila-allocycly-chromosome evolution-male sterility-dosage compensation). Proc Natl Acad Sci USA 1972;69:182–186.
76 Wu CI, Xu EY: Sexual antagonism and X inactivation: the SAXI hypothesis. Trends Genet 2003;19: 243–247.
77 Betran E, Thornton K, Long M: Retroposed new genes out of the X in *Drosophila*. Genome Res 2002;12:1854–1859.
78 Vinckenbosch N, Dupanloup I, Kaessmann H: Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA 2006;103:3220–3225.
79 Betran E, Emerson JJ, Kaessmann H, Long M: Sex chromosomes and male functions: where do new genes go? Cell Cycle 2004;3:873–875.
80 Khil PP, Oliver B, Camerini-Otero RD: X for intersection: retrotransposition both on and off the X chromosome is more frequent. Trends Genet 2005;21:3–7.
81 Emerson JJ, Kaessmann H, Betran E, Long M: Extensive gene traffic on the mammalian X chromosome. Science 2004;303:537–540.
82 Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H: Emergence of young human genes after a burst of retroposition in primates. PLoS Biol 2005;3:1970–1979.

Nadia D. Singh
371 Serra Mall
Stanford, CA 94305 (USA)
Tel. +1-607-254-4839, Fax +1-607-255-6249, E-Mail nds25@cornell.edu

Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 119–130

Amino Acid Repeats and the Structure and Evolution of Proteins

M.M. Albà[a], *P. Tompa*[b], *R.A. Veitia*[c]

[a]Research Unit on Biomedical Informatics, Catalan Institution for Research and Advanced Studies, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Barcelona, Spain; [b]Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest, Hungary; [c]INSERM E21, IFR Alfred Jost, Hôpital Cochin and University Paris VII, Paris, France

Abstract

Many proteins have repeats or runs of single amino acids. The pathogenicity of some repeat expansions has fueled proteomic, genomic and structural explorations of homopolymeric runs not only in human but in a wide variety of other organisms. Other types of amino acid repetitive structures exhibit more complex patterns than homopeptides. Irrespective of their precise organization, repetitive sequences are defined as low complexity or simple sequences, as one or a few residues are particularly abundant. Prokaryotes show a relatively low frequency of simple sequences compared to eukaryotes. In the latter the percentage of proteins containing homopolymeric runs varies greatly from one group to another. For instance, within vertebrates, amino acid repeat frequency is much higher in mammals than in amphibians, birds or fishes. For some repeats, this is correlated with the GC-richness of the regions containing the corresponding genes. Homopeptides tend to occur in disordered regions of transcription factors or developmental proteins. They can trigger the formation of protein aggregates, particularly in 'disease' proteins. Simple sequences seem to evolve more rapidly than the rest of the protein/gene and may have a functional impact. Therefore, they are good candidates to promote rapid evolutionary changes. All these diverse facets of homopolymeric runs are explored in this review.

Copyright © 2007 S. Karger AG, Basel

Many proteins have repeats of single amino acids, also known as homopolymeric runs [1]. This is the case of polyglutamines (polyQ), polyalanines (polyA), etc. Expansions of these monotonic tracts can be pathogenic. The most notable examples are polyQ expansions leading to neurodegenerative disorders [2]. Besides, there are an increasing number of proteins in which

polyA expansions beyond a critical threshold cause human disease [3]. For instance, expansions of the polyA domain of FOXL2, a protein involved in craniofacial and ovarian development, account for 30% of the reported intragenic mutations [4]. We have shown that this mutation induces the formation of intranuclear aggregates and a mislocalization of the protein [5]. The pathogenicity of these simple sequence expansions has stimulated proteomic and genomic explorations of homopolymeric runs in human and in a wide variety of other organisms.

Other types of amino acid repetitive structures exhibit more complex patterns than homopeptides. They include oligopeptide tandem repeats, cryptic repeats containing amino acid interruptions and prion-like Q/N-rich regions [6–9]. In this review, we focus on the genomics, proteomics, structural and functional aspects of homopolymeric runs and in general of such low complexity coding sequences.

Distribution of Coding Repeats and Homopeptides

Repeats of short DNA units (mono-, di-, trinucleotides, etc.) are abundant in genomic sequences. They are also known as microsatellites. Their predominant mode of mutation is thought to be replication slippage, which results in insertion or deletion (indels) of repeat units in the nascent DNA strand [10]. The mutation rate of indels is several orders of magnitude higher than that of point mutations [11], which explains the high microsatellite intra- and inter-specific divergence. Recombination may also contribute to the generation of repeat size variability [12]. Indeed, both mechanisms have been proposed to lead to length changes in polyA runs associated with human disease ([3], and references therein). Whereas in intergenic regions microsatellites composed of dinucleotides (i.e. AC, GT, etc.) are the most abundant ones, in coding sequences trinucleotide and hexanucleotide repeats predominate [13, 14] because they do not disrupt the open reading frame. The tendency of different trinucleotides to form microsatellite structures varies greatly within and between species [14].

The consequence of trinucleotide repeat expansion in coding regions is the formation of homopolymeric tracts in protein sequences, usually formed by hydrophilic amino acids [15–17]. Given the degeneracy of the genetic code, homopolymeric runs in proteins are not necessarily encoded by homogeneous trinucleotide repeats but may be encoded by a mixture of synonymous codons. Slippage activity, however, results in an excess of pure codon runs, particularly among evolutionarily non-conserved repeats [15, 18]. Prokaryotes show a relatively low frequency of homopeptides [17, 19], about one order of magnitude lower than eukaryotes (table 1). In eukaryotes, the percentage of proteins with

Table 1. Distribution of amino acid repeats and simple sequences in different species. Protein sequence datasets from complete genomes were obtained from Ensembl (28 February 2006) except for *E. coli*, *B. subtilis*, *A. fulgidus*, *M. jannaaschii* and *A. thaliana*, which were obtained from Cogent (31 March 2006)

	Species	No. of proteins	Average RSF[a]	Proteins with RSF >1 (%)[b]	Tandem AA repeats of ≥5 residues[c]	Most abundant AA in repeats of ≥5 residues	Tandem AA repeats of ≥10 residues[c]	Most abundant AA in repeats of ≥10 residues
Bacteria	*E. coli*	4,290	1.04	47	82 (0.019)	L(29), A(21), G(8)	0 (0.0)	–
	B. subtilis	4,093	1.04	46	84 (0.020)	L(19), S(18), A(15)	0 (0.0)	–
Archaea	*A. fulgidus*	2,409	1.03	46	50 (0.020)	E(11), V(9), L/P(6)	2 (0.001)	G(1), T(1)
	M. jannaaschii	1,773	1.04	43	27 (0.015)	E(7), K(6), L(3)	0 (0.0)	–
Eukarya	*S. cerevisiae*	6,680	1.16	65	1,246 (0.186)	S(298), Q(231), N(139)	144 (0.021)	Q(58), N(28), S(19)
	A. thaliana	25,761	1.25	73	6,694 (0.259)	S(1,746), G(832), E(798)	306 (0.011)	S(67), G(47), E(43)
Metazoa	*C. elegans*	26,032	1.20	66	5,569 (0.214)	S(804), Q(753), T(729)	181 (0.007)	S(42), T(41), Q(26)
	D. melanogaster	19,369	1.28	72	13,775 (0.711)	Q(4,299), A(2,094), S(1,606)	1,435 (0.074)	Q(700), A(293), G(111)
Chordata	*C. intestinalis*	20,000	1.15	62	2,209 (0.110)	E(283), S(263), P(249)	94 (0.004)	Q(28), S(15), T(15)
	F. rubripes	22,102	1.17	69	3,200 (0.145)	S(674), P(518), E(364)	50 (0.002)	S(16), P(12), A(9)
	G. gallus	24,268	1.19	68	2,986 (0.123)	S(588), E(519), P(434)	83 (0.003)	S(21), E(21), P(21)
Mammalia	*M. domestica*	26,943	1.14	65	4,143 (0.153)	P(747), S(661), E(609)	109 (0.004)	A(28), S(19), E(19)
	M. musculus	36,471	1.14	64	7,601 (0.208)	P(1,247), E(1,235), S(982)	783 (0.021)	E(193), Q(131), P(127)
	H. sapiens	33,869	1.17	68	8,259 (0.244)	E(1,420), P(1,325), A(1,130)	708 (0.021)	E(145), A(122), Q(119)

[a]RSF: relative simplicity factor. Values >1 indicate that the amount of simple sequence in the protein is higher than in random sequences. The program SIMPLE [20], with parameters window size = 2 and 100 randomizations shuffling amino acids, was employed to calculate RSF.

[b]$p < 0.05$.

[c]The number of amino acid repeats with 5 or more residues, or 10 or more residues, was obtained using an in-house Perl script (M.M. Albà). In parenthesis the number of repeats divided by the number of proteins analyzed.

at least one homopeptide of 5 or more residues is around 13% in *Caenorhabditis elegans*, 15% in *Saccharomyces cerevisiae*, 20% in *Homo sapiens* and *Arabidopsis thaliana*, and 27% in *Drosophila melanogaster* [1]. Within vertebrates, homopeptide frequency is much higher in mammals than in amphibians, birds or fishes [19]. In Chordata there is a predominance of polyE, polyP and polyS among repeats of 5 residues or longer, and of polyA and polyQ among the longest repeats. As previously noted [1], polyQ repeats stand out in *D. melanogaster*, with as many as 4,299 repeats in the fly proteome (table 1). In the plant *A. thaliana*, polyS repeats are the most abundant ones. PolyN is abundant in yeast [15] and in some non-vertebrate organisms such as insects and nematodes [19], but not in vertebrates. Table 1 shows the average Relative Simplicity Factor (RSF) of different prokaryotic and eukaryotic organisms [20]. RSF values above 1 indicate an excess of simple sequence in a protein with respect to the random expectation. Bacteria and Archaea show average RSF values only slightly over 1 (1.03–1.04), much lower than those of eukaryotes (e.g. *S. cerevisiae* 1.16, *D. melanogaster* 1.28, *H. sapiens* 1.17).

Relationship between GC Content and Repeat Content

Several studies have pointed to the existence of a relationship between GC content and the frequency of homopeptides in mammalian genes. The first reports dealt with class III POU transcription factors [21], and showed a significant positive correlation between the GC content at the third codon position (GC3) and runs of polyA, polyG and polyP. It was then suggested that nucleotide compositional constraints increasing the GC content of coding sequences would facilitate the formation of repeats in GC-rich genes, a hypothesis supported by subsequent studies [22, 23]. Recombination seems to be a key determinant of local GC-enrichment and is associated with high density of minisatellites [24]. Therefore, it would not be surprising to find a correlation between recombination hot spots and the presence of coding (and non-coding) microsatellites at a genomic scale. As noted above, sequence misalignment during recombination may also play a role in length variation of coding repeats.

Coding Repeats and Protein Function

Early studies documented the presence of homopeptides in particular types of proteins, such as transcription factors [25] and developmental proteins

[26, 27]. More recent genome-wide studies on *S. cerevisiae* showed a significant enrichment of transcription factors among proteins containing polyQ, polyN or polyD [15]. Metabolic functions, on the other hand, were under-represented in homopeptide-containing proteins [28]. A comprehensive analysis of *H. sapiens* and *D. melanogaster* genomes confirmed the noted abundance of transcription factors and developmental proteins among proteins containing tandem repeats [1]. In mammalian proteins, polyA was over-represented in proteins involved in RNA- and DNA-binding, and polyL in transmembrane receptors [23]. Besides, it was observed that many proteins containing homopeptides performed roles related to the assembly of large multiprotein or protein/nucleic acid complexes [19]. This functional bias is related to the preference of tandemly repeated amino acids to occur in intrinsically unstructured proteins (IUPs), and the strong correlation of IUPs with functions of signal transduction and transcription regulation [29].

Structure of polyA and polyQ Repeats

Given the involvement of polyA and polyQ expansions in human disease, numerous structural studies have addressed these regions. Short polyA runs are rather disordered, fluctuating over a range of alternative conformations, including α-helix, left-handed polyproline II helix (PPII) and β-strand. For three A residues, a temperature-dependent mixture of PPII and extended β-strand conformations has been observed [30], whereas the PPII conformation becomes more stabilized with an increasing number of residues [31]. Molecular Dynamics (MD) simulations have suggested that four of the seven A residues in the Ac-X_2-A_7-O_2-NH_2 peptide have a conformational preference for β-strand or extended structure, whereas in the Ac-A_8-NH-Me peptide the largest cluster of conformations corresponds to a β-hairpin [32]. However, a polyA tract flanked by lysine residues in Ac-K-A_7-K-NH_2 was shown to exist predominantly in α-helical conformation [33, 34].

PolyQ stretches prefer a disordered state with some tendency to form β-structures depending on their sequential environment. A random coil structure was found in polyQ stretches of 9 or 17 residues flanked by sequences rich in A and K [35], or in a G-Q_{10}-G peptide inserted into the inhibitory loop of chymotrypsin inhibitor 2 [36]. Under slightly non-physiological, and possibly denaturing, conditions (i.e. low pH) a significant population of more ordered structures, such as β-hairpins and other types of β-structures were observed [37]. Thus, the conformation of polyQ regions is very sensitive to the molecular environment, with a noted tendency to transit from the random-coil conformation to various β-structures.

Structure of other Homopolymers and Nearly Perfect Repeats

As noted above, runs of E, P, S, R, D and K occur with high frequency in eukaryotes [1, 17]. Most data suggest that these repetitive stretches do not adopt a well-defined structure, but occur in disordered conformations. This has been demonstrated experimentally for polyE [38] and polyS regions [39], and also for the Q/N-rich domains of yeast prions, such as Sup35p [40]. Recent experiments have shown that disorder in these IUPs may locally correspond to the PPII structure [41]. This was also demonstrated for model peptides containing runs of S, E, D, Q and V [42]. The major functional implication of this finding might be that PPII plays a key role in protein-protein interactions [43], which may also be directly linked to its pathological involvement due to its tendency for homotypic interactions and the formation of amyloid.

Formation of Aggregates by Homopeptides

Homopolymeric amino acid runs and nearly perfect repeats have a high propensity to form aggregates, primarily noted in the case of polyQ, polyA and prion-like Q/N-rich sequences. The tendency to form aggregates is a result of a fine structural interplay between the amino acid run, the flanking region and environmental effects. As mentioned above, it has been noted that flanking residues may alter local preferences, leading to disordered or repetitive conformations [44]. The fact, that in different polyA diseases the critical pathological length of the repeat region shows significant variation, may also be explained by the distinct effects of the flanking regions [3]. A similar sensitive behavior has been noted for polyQ [45] and polyE sequences [38]. For example, polyQ stretches of 22 and 41 residues fused to glutathion S-transferase (GST) prefer the random coil conformation [45], whereas short polyQ regions below the pathological length ($<$40) inserted into sperm whale myoglobin form intramolecular β-sheets buried in the core of the protein [46]. Thus, fusion to a 'solubilizer' domain (such as GST) often keeps the polyQ region in a soluble state [45], while aggregation is induced by cleavage of the GST domain.

Besides their direct participation in aggregate formation, repeat regions may also affect their sequential neighborhood. Spectroscopic studies of the wild-type Q_{27}-ataxin-3 and the pathological Q_{78}-ataxin-3 indicated a high α-helical content (35% α-helix) for the shorter stretch, and the loss of helical structure and a gain of random coil for Q_{78}-ataxin-3 [47]. Even more subtle effects are seen in the case of huntingtin, which contains a polyP region C-terminal to the characteristic polyQ stretch. This polyP significantly inhibits

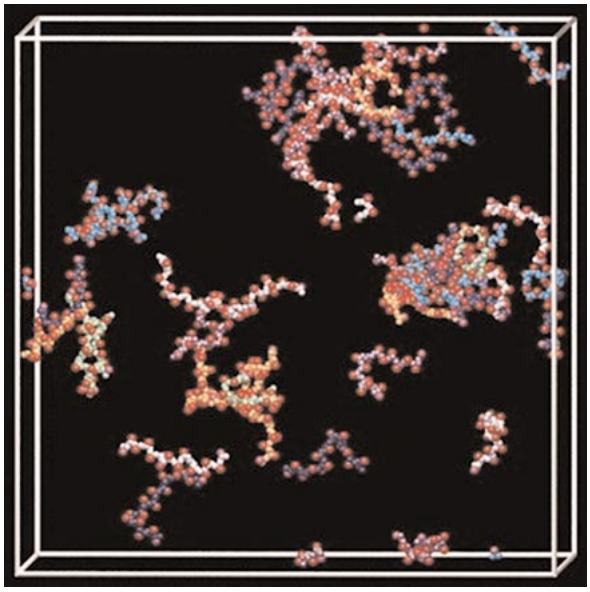


Fig. 1. Simulated structure of amyloid. Molecular dynamics simulation of 48 Ala_{14} peptides shows that a transient critical step towards amyloid formation is the assembly of loose aggregates of rather disordered peptides. At a later step (not shown), the aggregates transform to the typical cross-β structure via partial disassembly and seeding by smaller β-units. Reproduced from reference [50] with permission from Proc. Natl. Acad. Sci. USA.

aggregate formation by the polyQ region, but when it is moved to the N-terminal side of polyQ, its effect is completely abolished [48].

Amyloid formation (i.e. a kind of aggregation process) has been observed in the case of many proteins, which have practically nothing in common in terms of either sequence or structure [49]. The unifying theme in all their conversions is that the protein in the transition state has to attain a partially structured state [49]. For ordered proteins, amyloidosis preferentially occurs under conditions that destabilize the native structure, whereas for IUPs, amyloid formation is promoted by conditions, which favor partially ordered species. For instance, the Q/N-rich domain of the Sup35p yeast prion converts to amyloid via a transient aggregated state [40] and structural reorganization occurs within this collapsed, but still sufficiently flexible state. For the homopolymers covered in this review, MD simulations provide some insight into the critical steps of the transition to an aggregate state [50]. In an ensemble of oligoA peptides, transition to the aggregated state begins with the formation of amorphous aggregates of several rather disordered molecules (fig. 1). The aggregate then falls apart to smaller pieces, some of which have the distinctive parallel β-like

arrangement of chains. Isolated chains are then very effectively captured and converted to this orderly structure by these seeds.

Structure of Amyloid

No high-resolution structural image of amyloid has been available for a long time. Structural models have been built upon the information provided by fiber diffraction and other low-resolution techniques, which all agree in the cross-β structure of the chain and tight intermolecular packing within the amyloid fiber. The most influential structural model for polyQ amyloids has been the 'polar zipper', which assumes that adjacent antiparallel β-strands run perpendicular to the fiber axis, and are held together by a network of H-bonds between main-chain and side-chain amide bonds [37]. Recent observations on Q/N amyloids are more compatible with parallel strands in the sheet, whereas in the case of polyA sequences the arrangement of chains is envisaged as a result of close packing of adjacent methyl groups, which results in stabilizing hydrophobic interactions and tight interdigitation of side-chains of adjacent sheets [50]. X-ray crystallography of the GNNQQNY heptapeptide derived from the amyloid domain of Sup35p is now available [51]. The structure in the crystal, termed 'steric zipper', is considered compatible with the amyloid structure itself. Its common relevance is warranted by MD simulations of the structure of polyA amyloids, which also show a similar mode of tight packing excluding water [50].

Coding Repeats and Evolution

Numerous studies indicate that homopolymeric runs and simple sequences tend to be poorly conserved across species, which may be related to relaxed selection pressure but also to taxon-specific functional diversification. Faux et al. reported that out of 20 prokaryote families of proteins containing repeats and with eukaryotic homologues, only three had conserved similar repeat tracts in the eukaryotic proteins, indicating a low level of conservation across long phylogenetic distances [19]. Analysis of several vertebrate class III POU transcription factors showed that most of the polyA, polyG and polyP runs present in the mammalian proteins were absent in the amphibian and fish homologues [21]. A similar lack of conservation of polyA runs was observed in HOXA13 proteins [52]. Phylogenetic analysis of the transcription factors HOX, GATA and EVX has shown that different polyAs arose independently in different lineages [53]. The same work showed that, in general, polyAs were conserved among mammals but were rarer and shorter in *G. gallus* and *D. rerio*.

Analysis of simple sequences in yeast has shown that these regions evolve more rapidly than the remainder of the protein sequence [54]. Simple sequences across orthologous protein families are in general poorly conserved [8]. Interestingly, in humans, repetitive structures of amino acids are more common in recent proteins than in ancient proteins (i.e. those with homologues in *E. coli* [55]). In agreement, by comparing mammalian-specific, deuterostome-specific, metazoan and even older proteins, Albà and Castresana observed that the fraction of a protein occupied by simple sequences showed an inverse correlation with the 'age' of the protein [56]. This data suggests that these regions could play an important role in enhancing functional flexibility.

Mutations in *cis*-regulatory elements of developmental genes have been thought to be the basis of organismal diversity. However, length variations of homopolymeric runs in developmental genes may also be associated with profound changes. For instance, Galant and Carroll [57] found that the polyA run in the Ultrabithorax (Ubx) protein was conserved among insect Ubx but absent in orthologs from other arthropods and onychophorans. The presence of this polyA domain in insects is thought to have led to suppression of abdominal limbs [57]. Human HOX and other developmental proteins carry polyA and other non-conserved repeats which may underlie changes in their activity [53]. This may to some extent explain differences in a macro-evolutionary scale (i.e. among vertebrates).

Coding repeats may also play relevant roles in microevolution. They show high polymorphism levels, especially when repeats are long and encoded by pure or nearly pure triplet runs [53, 58]. In an interesting study, Fondon and Garner have shown that length variations of coding repeats in developmental genes in dog breeds are associated with rapid morphological changes [59]. Not surprisingly, they showed a tendency to repeat purity in dogs, suggesting that the various alleles resulted from recent changes driven by polymerase slippage. For example, the authors could associate the bilateral rear first digit polydactyly found in the Great Pyrenees dog with a homozygous deletion within a P/Q repeat in the *ALX4* gene. Fondon and Garner also found a positive correlation between craniofacial morphological parameters and the ratio of polyQ/polyA of RUNX2, a key regulator of osteoblast differentiation implicated in cleidocranial dysplasia (CCD) in man. This correlation was attributed to a possible modulation of its activity via a transcriptional activating polyQ domain in opposition to a polyA repressive domain. Interestingly, a mild form of familial human CCD results from a 10-Ala-expansion of the polyA tract of RUNX2. Thus, it is also conceivable (for RUNX2 and other polyA factors) that different polyA alleles would induce a differential protein aggregation, thus regulating the pool of soluble factors available for transcriptional activity. Recent studies have shown that polyA expansions not leading to any visible aggregation can also be associated

with a decrease of transcriptional activity. In these cases, there may be a population of different conformers with distinct transcriptional activities, also affected by chaperones and post-translational modifications [60].

As mentioned above, there is a close link between regional DNA properties and the presence of certain types of coding repeats. The connection is particularly clear for GC-richness and the presence of polyA, G and P. An interesting perspective is that the stronger isochoric patterning in mammals compared to other vertebrates, with more marked GC-rich and GC-poor genomic regions, has influenced the number of homopeptidic runs in this taxonomic group. Since these runs may have an impact on protein structure and function, genomic constraints will indirectly leave morphological and physiological traces. It would be interesting to assess if this is a more general phenomenon. As previously said, there is mounting evidence involving polyA and polyQ expansions in human disease. Is that all, or just the tip of the iceberg?

References

1 Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ: Amino acid runs in eukaryotic proteomes and disease associations. Proc Natl Acad Sci USA 2002;99:333–338.
2 Ross CA: When more is less: pathogenesis of glutamine repeat neurodegenerative diseases. Neuron 1995;15:493–496.
3 Brown LY, Brown SA: Alanine tracts: the expanding story of human illness and trinucleotide repeats. Trends Genet 2004;20:51–58.
4 De Baere E, Dixon MJ, Small KW, Jabs EW, Leroy BP, et al: Spectrum of *FOXL2* gene mutations in blepharophimosis-ptosis-epicanthus inversus (BPES) families demonstrates a genotype–phenotype correlation. Hum Mol Genet 2001;10:1591–1600.
5 Caburet S, Demarez A, Moumne L, Fellous M, De Baere E, Veitia RA: A recurrent polyalanine expansion in the transcription factor FOXL2 induces extensive nuclear and cytoplasmic protein aggregation. J Med Genet 2004;41:932–936.
6 Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D: A census of protein repeats. J Mol Biol 1999;293:151–160.
7 Tompa P: Intrinsically unstructured proteins evolve by repeat expansion. BioEssays 2003;25:847–855.
8 Sim KL, Creamer TP: Protein simple sequence conservation. Proteins 2004;54:629–638.
9 Hancock JM, Simon M: Simple sequence repeats in proteins and their significance for network evolution. Gene 2005;345:113–118.
10 Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN: Complex gene rearrangements caused by serial replication slippage. Hum Mutat 2005;26:125–134.
11 Ellegren H: Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet 2000;16:551–558.
12 Li Y, Korol AB, Fahima T, Nevo E: Microsatellites within genes: structure, function and evolution. Mol Biol Evol 2004;21:991–1007.
13 Stallings RL: Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases. Genomics 1994;21:116–121.
14 Tóth G, Gáspári Z, Jurka J: Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 2000;7:967–981.
15 Albà MM, Santibáñez-Koref MF, Hancock JM: Amino acid reiterations in yeast are over-represented in particular classes of proteins and show evidence of a slippage-like mutational process. J Mol Evol 1999;49:789–797.

16 Green H, Wang N: Codon reiteration and the evolution of proteins. Proc Natl Acad Sci USA 1994;91:4298–4302.
17 Karlin S, Burge C: Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. Proc Natl Acad Sci USA 1996;93:1560–1565.
18 Albà MM, Santibáñez-Koref MF, Hancock JM: Conservation of polyglutamine tract size between mouse and human depends on codon interruption. Mol Biol Evol 1999;16:1641–1644.
19 Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, et al: Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res 2005;15:537–551.
20 Albà MM, Laskowski RA, Hancock JM: Detecting cryptically simple protein sequences using the SIMPLE algorithm. Bioinformatics 2002;18:672–678.
21 Nakachi Y, Hayakawa T, Oota H, Sumiyama K, Wang L, Ueda S: Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. Mol Biol Evol 1997;14:1042–1049.
22 Cocquet J, De Baere E, Caburet S, Veitia RA: Compositional biases and polyalanine runs in humans. Genetics 2003;165:1613–1617.
23 Albà MM, Guigó R: Comparative analysis of amino acid repeats in rodents and humans. Genome Res 2004;14:549–554.
24 Montoya-Burgos JI, Boursot P, Galtier N: Recombination explains isochores in mammalian genomes. Trends Genet 2003;19:128–130.
25 Gerber HP, Seipel K, Georgiev O, Hofferer M, Hug M, et al: Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 1994;263:808–811.
26 Treier M, Pfeifle C, Tautz D: Comparison of the gap segmentation gene hunchback between *Drosophila melanogaster* and *Drosophila virilis* reveals novel modes of evolutionary change. EMBO J 1989;8:1517–1525.
27 Newfeld SJ, Smoller DA, Yedvobnick B: Interspecific comparison of the unusually repetitive *Drosophila* locus mastermind. J Mol Evol 1991;32:415–420.
28 Young ET, Sloan JS, van Riper K: Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. Genetics 2000;154:1053–1068.
29 Dunker AK, Lawson JD, Brown CJ, Romero P, Oh JS, et al: Intrinsically disordered protein. J Mol Graph Model 2001;19:26–59
30 Mu Y, Kosov DS, Stock G: Conformational dynamics of trialanine in water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments. J Phys Chem B 2003;107:5067–5073.
31 Schweitzer-Stenner R, Eker F, Griebenow K, Cao X, Nafie LA: The conformation of tetraalanine in water determined by polarized Raman, FT-IR, and VCD spectroscopy. J Am Chem Soc 2004;126:2768–2776.
32 Ramakrishnan V, Ranbhor R, Durani S: Existence of specific 'folds' in polyproline II ensembles of an 'unfolded' alanine peptide detected by molecular dynamics. J Am Chem Soc 2004;126: 16332–16333.
33 Chakrabartty A, Kortemme T, Baldwin RL: Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. Protein Sci 1994;3:843–852.
34 Blondelle SE, Forood B, Houghten RA, Perez-Paya E: Polyalanine-based peptides as models for self-associated beta-pleated-sheet complexes. Biochemistry 1997;36:8393–8400.
35 Altschuler EL, Hud NV, Mazrimas JA, Rupp B: Random coil conformation for extended polyglutamine stretches in aqueous soluble monomeric peptides. J Pept Res 1997;50:73–75.
36 Gordon-Smith DJ, Carbajo RJ, Stott K, Neuhaus D: Solution studies of chymotrypsin inhibitor-2 glutamine insertion mutants show no interglutamine interactions. Biochem Biophys Res Commun 2001;280:855–860.
37 Perutz MF, Johnson T, Suzuki M, Finch JT: Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. Proc Natl Acad Sci USA 1994;911:5355–5358.
38 Idiris A, Alam MT, Ikai A: Spring mechanics of alpha-helical polypeptide. Protein Eng 2000;13: 763–770.
39 Howard MB, Ekborg NA, Taylor LE, Hutcheson SW, Weiner RM: Identification and analysis of polyserine linker domains in prokaryotic proteins with emphasis on the marine bacterium *Microbulbifer degradans*. Protein Sci 2004;13:1422–1425.

40 Krishnan R, Lindquist SL: Structural insights into a yeast prion illuminate nucleation and strain diversity. Nature 2005;435:765–772.
41 Blanch EW, Kasarda DD, Hecht L, Nielsen K, Barron LD: New insight into the solution structures of wheat gluten proteins from Raman optical activity. Biochemistry 2003;42:5665–5673.
42 Shi Z, Chen K, Liu Z, Sosnick TR, Kallenbach NR: PII structure in the model peptides for unfolded proteins: studies on ubiquitin fragments and several alanine-rich peptides containing QQQ, SSS, FFF, and VVV. Proteins 2006;63:312–321.
43 Williamson MP: The structure and function of proline-rich regions in proteins. Biochem J 1994;297: 249–260.
44 Chen K, Liu Z, Zhou C, Shi Z, Kallenbach NR: Neighbor effect on PPII conformation in alanine peptides. J Am Chem Soc 2005;127:10146–10147.
45 Masino L, Kelly G, Leonard K, Trottier Y, Pastore A: Solution structure of polyglutamine tracts in GST-polyglutamine fusion proteins. FEBS Lett 2002;513:267–272.
46 Tanaka M, Morishima I, Akagi T, Hashikawa T, Nukina N: Intra- and intermolecular beta-pleated sheet formation in glutamine-repeat inserted myoglobin as a model for polyglutamine diseases. J Biol Chem 2001;276:45470–45475.
47 Bevivino AE, Loll PJ: An expanded glutamine repeat destabilizes native ataxin-3 structure and mediates formation of parallel beta-fibrils. Proc Natl Acad Sci USA 2001;98:11955–11960.
48 Bhattacharyya A, Thakur AK, Chellgren VM, Thiagarajan G, Williams AD, et al: Oligoproline effects on polyglutamine conformation and aggregation. J Mol Biol 2006;355:524–535.
49 Uversky VN, Fink AL: Conformational constraints for amyloid fibrillation: the importance of being unfolded. Biochim Biophys Acta 2004;1698:131–153.
50 Nguyen HD, Hall CK: Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. Proc Natl Acad Sci USA 2004;101:16180–16185.
51 Nelson R, Sawaya MR, Balbirnie M, Madsen AO, Riekel C, et al: Structure of the cross-beta spine of amyloid-like fibrils. Nature 2005;435:773–778.
52 Mortlock DP, Sateesh P, Innis JW: Evolution of N-terminal sequences of the vertebrate HOXA13 protein. Mamm Genome 2000;11:151–158.
53 Lavoie H, Debeane F, Trinh QD, Turcotte JF, Corbeil-Girard LP, et al: Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains. Hum Mol Genet 2003;12:2967–2979.
54 Huntley M, Golding GB: Evolution of simple sequences in proteins. J Mol Evol 2000;51:131–140.
55 Nishizawa M, Nishizawa K: Local-scale repetitiveness in amino acid use in eukaryote protein sequences: a genomic factor in protein evolution. Proteins 1999;37:284–292.
56 Albà MM, Castresana J: Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol 2005;22:598–606.
57 Galant R, Carroll SB: Evolution of a transcriptional repression domain in an insect Hox protein. Nature 2002;415:910–913.
58 Wren JD, Forgacs E, Fondon JW 3rd, Pertsemlidis A, Cheng SY, et al: Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. Am J Hum Genet 2000;67:345–356.
59 Fondon JW 3rd, Garner HR: Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci USA 2004;101:18058–18063.
60 Caburet S, Cocquet J, Vaiman D, Veitia RA: Coding repeats and evolutionary ‘agility’. Bioessays 2005;27:581–587.

Reiner A. Veitia
INSERM E21
IFR Alfred Jost, Hôpital Cochin, Pavillon Baudelocque
123 Bd de Port Royal, 75014 Paris (France)
Tel. +33 1444 12317, Fax +33 1444 12302, E-Mail veitia@cochin.inserm.fr

Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 131–146

Origination of Chimeric Genes through DNA-Level Recombination

J.R. Arguello[a], *C. Fan*[a], *W. Wang*[b], *M. Long*[a]

[a]The University of Chicago, Department of Ecology and Evolution, Zoology 301E, Chicago, Ill., USA; [b]CAS-Max Planck Junior Research Group, Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China

Abstract

Comparative genomics is rapidly bringing to light the manifold differences that exist within and between species on the molecular level. Of fundamental interest are the absolute and relative amounts of the genome dedicated to protein coding regions. Results thus far have shown surprising variation on both the polymorphism and divergence levels. As a result, there has been an increase in efforts aimed to characterize the underlying genetic mechanisms and evolutionary forces that continue to alter genomic architecture. In this review we discuss the formation of chimeric genes generated at the DNA level. While the formation of chimeric genes has been shown to be an important way in which coding regions of the genome evolve, many of the detailed studies have been limited to chimeric genes formed through retroposition events (through an RNA intermediate step). Here we provide a short review of the reported mechanisms that have been identified for chimeric gene formations, excluding retroposition-related cases, and discuss several of the evolutionary analyses carried out on them. We emphasize the utility chimeric genes provide for the study of novel function. We also emphasize the importance of studying chimeric genes that are evolutionarily young.

Copyright © 2007 S. Karger AG, Basel

Fundamental to an understanding of biological diversity are questions pertaining to the acquisition of new functions. As a result, the origin of evolutionary novelty has long captured the imagination of evolutionary biologists and naturalists. On the molecular scale, chimeric genes – by which we are referring to gene structures (both coding and noncoding) that have been derived from multiple parental loci – provide a useful system for studying the origin of new functions. The first reason for this is that chimeric genes are unlikely to have

redundant functions. This is in contrast to classical views of gene duplication in which theoretical models assume that an exact duplicate copy possessing redundant functions is initially produced [1, 2]. The second reason is that recent findings suggest chimeric proteins may arise at unexpectedly high rates [3–8], and, because of this, there is particular excitement over a growing number of reports on evolutionarily young cases. The identification of young cases represents an important advance for the field because it provides the opportunity to examine the early evolutionary steps involved in the acquisition of new functions. The third reason is that chimeric structures aid in the resolution of ancestor-offspring relationships. The ability to distinguish the derived copy from the ancestral copy becomes a substantial issue when trying to understand evolutionary forces that may differ between paralogs.

Considerable insight into the evolution of chimeric genes has recently been gained through the study of retroposed genes. Both genome-wide analyses and individual case studies have demonstrated that numerous new genes with novel functions have been generated as a result of initial retropositions [3, 4, 7, 9–14]. However, less is known about the origin of chimeric genes through recombination events at the DNA level. In this chapter our goal is to review data regarding this less understood phenomenon. In doing so we will omit instances of chimeric formations that have occurred through an RNA intermediate step, but for those interested we refer to several reviews [15–17]. This chapter is further narrowed by our treatment of sequence-level studies and not biochemical and molecular biology studies. Both are relevant, and have made major contributions to our understanding of the underlying genetic mechanisms, but they are beyond the scope of this chapter.

We have organized this review into two sections. We first introduce the genetic mechanisms that have been shown to be capable of generating chimeric proteins on the DNA level. Next, we focus on illustrious examples that provide evidence for each of these mechanisms. Nonhomologous recombination (NHR) will comprise the majority of this second section, and we have divided these data, and the corresponding experimental approaches, into ‘evolutionarily ancient’ and ‘evolutionarily recent’ examples. In the ‘evolutionarily recent’ section we present a chimeric protein, *Hun*, which our group recently reported on [18]. Throughout the chapter, we highlight the utility of young chimeric proteins as a system for studying novel functions.

Molecular Mechanisms Leading to Chimeric Genes

At the level of DNA, several molecular mechanisms have been observed to recombine different genic and nongenic regions to generate chimeric genes

structures. Nonhomologous recombination was the mechanism first proposed [19]. In this early model, known as exon-shuffling, NHR within introns leads to novel combinations of exonic regions. More recently, models for chimeric gene origination have become more numerous and more flexible as the number of recognized mechanisms giving rise to them continues to increase.

Nonhomologous Recombination

NHR, also called illegitimate recombination, refers to recombination events that occur without reliance on extensive stretches of identity, and instead occur between regions with little or no identity [20–22]. Both NHR and homologous recombination (HR) pathways occur efficiently, and likely in an overlapping manner, to repair double stranded breaks (DSB) [23–25]. HR usually results in accurate strand repair [23, 25, 26] while NHR results in imperfect repair, causing duplications, insertions, and deletions. The involvement of gene regions in any of these latter events could potentially result in novel chimeric genes.

Non-allelic Homologous Recombination

When strand repair occurs through HR, or if there is mispairing during synapsis, there is the potential for different low copy repeats (LCRs) to recombine [27, 28]. This is referred to as non-allelic homologous recombination (NAHR). LCRs are generally short blocks (1–400 kb) of duplicated DNA which share considerable sequence identity [28–30]. NAHR events result in numerous rearrangements including duplications, inversions, deletions and translocations (fig. 1). Similar to NHR, if the breakpoints of these events involve gene regions, chimeric genes may result.

Transposable Elements as 'Fragment Joiners'

Transposable elements (TEs), being a type of LCR, can be involved in the origin of chimeric genes through the facilitation of NAHR (above). However, they also have the capability of recombining short DNA sequences through their inherent biochemical processes. Currently, this has been recognized with two plant TEs, Pack-MULEs and *Helitrons*. Though the mechanism is still uncertain, Pack-MULEs can recruit small chromosome fragments and combine with other genomic regions through their own movements to form chimeric gene structures [31]. It is thought that ectopic gene-conversion across a nicked cruciform structure may play a role in this recruitment process [32] (fig. 2). *Helitrons*, which are helicase-bearing transposable elements, are likewise capable of shuffling genomic regions. Again, the mechanism has not been worked out for these TEs, but they are capable of transporting replicated elements to target sequence and replacing it with its own DNA. This has been referred to as a rolling circle transposition mechanism [33, 34].

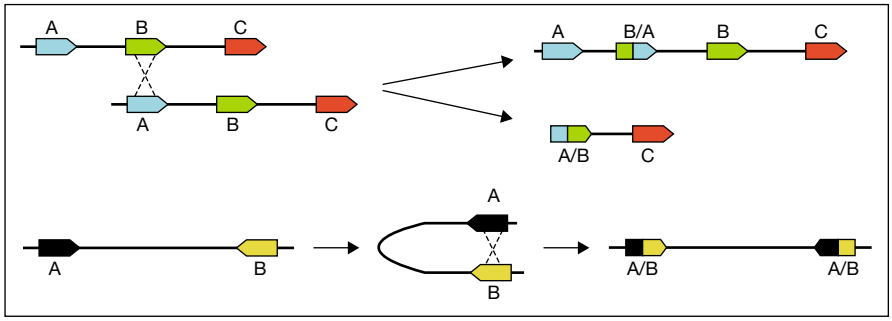


Fig. 1. Models representing non-allelic homologous recombination. The top panel depicts a recombination event between low-copy repeats (pointed boxes A–C). The products of this event are a duplicated (top right) and a deficient (bottom right) chromosome region. Note that the recombination event can occur between homologous or nonhomologous chromosomes. If the event occurs between nonhomologous chromosomes, a translocation would be produced. The bottom panel depicts a recombination event between two low-copy repeats that exist on the same chromosome, but results in an inverted configuration. Such non-allelic recombination events can occur through hairpin structures.

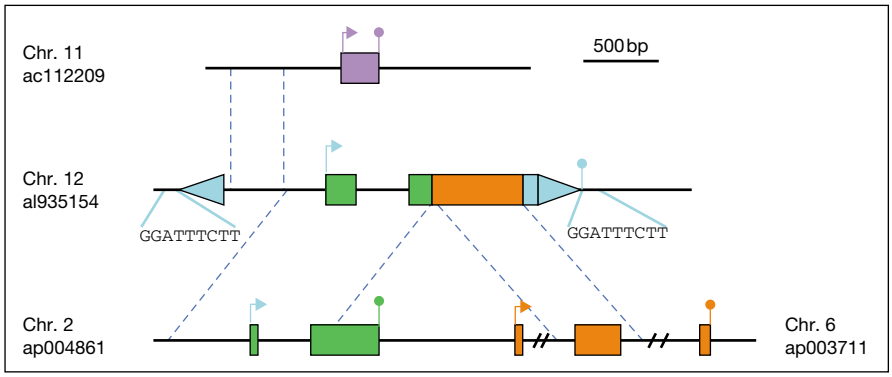


Fig. 2. Chimeric gene formation involving a Pack-MULE. The model provides an example of a novel chimeric gene (al935154 on chr. 12) that was created by a Pack-MULE containing gene fragments from three genomic loci, including both introns and exons (ac112209 at chr.11, ap004861 at chr. 2, and a putative bHLH transcription factor at chr. 6). The tandem inverted repeat is noted by the purple sequence and the start and stop codons are marked for each gene with an arrow and dot, respectively. Homologous fragments are indicated by dash lines (figure modified from [31]).

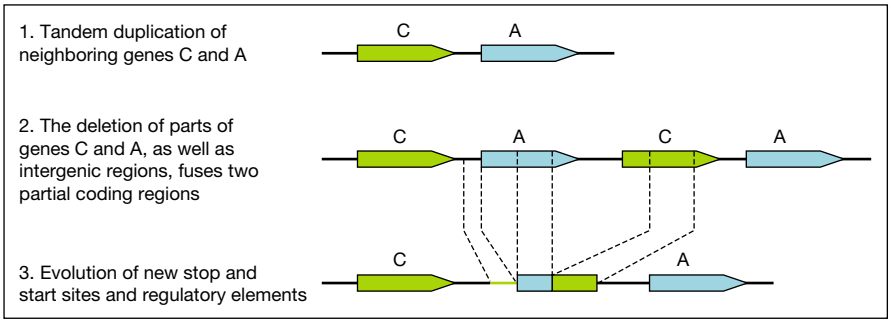

1. Tandem duplication of neighboring genes C and A
C
A
2. The deletion of parts of genes C and A, as well as intergenic regions, fuses two partial coding regions
C
A
C
A
3. Evolution of new stop and start sites and regulatory elements
C
A

Fig. 3. Example of a chimeric gene formed by gene fusion. This model is a simplified version of that found in [35]. In it, a gene pair (C and A) is duplicated in tandem. The duplication is followed by deletions that combine the remaining exonic regions of the two middle genes (A and C). Later evolutionary events include the recruitment of regulatory elements and the establishment of new start and stop codons if they were deleted.

Gene Fusions

It has been shown that two distinct genes or gene regions in adjacent genomic positions can be fused to form a single chimeric gene. A requirement for generating a chimeric gene in this way is to delete or skip the stop codon in the upstream gene. In prokaryotic genomes, the stop codon can be eliminated through nucleotide insertions, as shown in a fusion experiment using *E. coli* tryptophan synthetase alpha and beta polypeptides [35]. In eukaryotic genomes, two naturally occurring examples of gene fusion events have been identified and have been shown to be the result of two different molecular processes. One instance resulted from multiple deletions in a region between two neighboring genes, while the other resulted from unusual alternative splicing across two genes ([36, 37]; fig. 3 and below).

Evidence for Chimeric Proteins

Nonhomologous Recombination: Ancient Chimeras

In detecting ancient chimeric genes, two general methods have been used. The first is protein sequence comparisons in which similarity searches are carried out between protein regions from different protein families. Second, statistical analyses have been used to detect signals left over from the shuffling processes, for example the phase of introns (the positions of introns within and between codons; see below).

The first documented chimeric genes created by exon shuffling were the human tissue plasminogen activator (TPA) [38, 39] and the low-density lipoprotein (LDL) receptor protein in humans [39, 40]. TPA, which is necessary for the conversion of plasminogen into its fibrin dissolving active form, was found to be composed of domains which share significant similarity with uroinase, epidermal growth factor, and fibronectin domains. The LDL receptor, which is a cell-surface protein that mediates the endocytosis of low density lipoprotein (LDL), has eighteen exons. Thirteen of the eighteen exons share significant similarity with other proteins such as the C9 component factor, the EGF precursor, and blood clotting factors (IX, X, and C). Following these studies, Patthy did intensive sequence comparisons which revealed more than 300 gene families that contain mosaic domain structures [41–44]. Included in Patthy's dataset are proteins found in the coagulation cascade of mammals and fish, indicating an ancient generation dating back approximately 450 million years ago [41, 42]. Also notable is the fact that many intracellular proteins involved in signaling pathways are present in humans, worm, and yeast which share a remote common ancestor (1.46 BYA) [42, 43].

While intriguing, the chimeric structures alone are not sufficient evidence for exon shuffling as originally proposed, in which introns play an essential role in the recombination process [19]. Additional support for the role of exon shuffling in the origination of these chimeric genes comes from further analyses demonstrating that the recruited domains are flanked by introns of identical phases (1, 1) (i.e. symmetric exons). This is a hallmark of exon shuffling because it maintains the reading frames of the new gene [44].

In both plants and animals, kinases provide canonical examples of exon shuffling, supported by both sequence identity and phase data [41]. These data indicate that exon shuffling events through DNA level recombination took place approximately 990 million years ago according to the estimate of divergence time for these organisms [45]. In plants, the receptor kinases possess functions equivalent to animal receptor tyrosine kinases [41], but it is unclear whether the low sequence similarity is due to independent originations or the long evolutionary time that separates animals and plants (1.58 billion years). Another insightful instance of exon shuffling is found in the origination of the cytochrome c1 precursor gene in potato. Cytochrome c1 is part of the mitochondrial respiratory chain and is found in most eukaryotes. In potato, it was found that the mitochondria-derived nuclear gene, cytochrom c1, recruited a target domain from GapC. This shuffling event resulted in a mitochondrial targeting function for the new protein [46].

Taking advantage of the available sequenced genomes, as well as recent methods used to define protein domains, several groups have taken a genome-wide approach to investigating the relationships between domains and introns. In a human genome study, Kaessman et al. [47] analyzed the effect of the position

of introns, within or outside domain boundaries, on the distribution of intron phases. They observed that the introns within the domain boundaries show significant intron phase correlations (for example the excess of (1,1) symmetric exons), however, the introns outside the boundaries (or within domains) do not show such correlations. These results reveal a role for exon shuffling in recombining protein domains to form chimeric genes. Similarly, in a study that included the genomes of human, mouse, rat, *Fugu*, zebrafish, *Drosophila*, mosquito, *C. briggsae*, and *C. elegans*, Liu and Grigoriev [48] found a significant correlation between exon borders and protein domain borders. Interestingly, the significance of this correlation increased as they moved from *C. briggsae* to humans. Moreover, they also found that most of the exon-correlating domains were bordered by symmetric introns, with 1-1 introns being the most frequent.

As a cautionary note we would like to point out that it has been shown that an intron-containing gene structure can also be transposed into a new genomic position by retroposition of the gene's antisense RNAs [5]. This can potentially recombine with preexisting exon-intron structures to form a new chimeric gene. Given recent findings that show a high proportion of genes in mammals and invertebrates have antisense RNAs [49, 50], an alternative retroposition model for some of the examples above cannot be discounted. To be able to exclude competing models of chimeric gene formation, recently evolved chimeric genes are often necessary.

Nonhomologous Recombination – Evolutionarily Recent Chimeras

Not all of the NHR events that result in chimeric gene structures fit within the exon-shuffling model. This has been most clearly demonstrated through studies of young cases in which, among other means, chimeric genes have formed through recombination events within exons as well as through the recruitment of previously nongenic DNA. Studies of recent NHR events leading to chimeric proteins primarily utilize breakpoint data obtained from case studies of naturally occurring NHR events [28–30, 51–54] or transfection-based experimental approaches [20–22, 24, 55–57]. Both methods present particular benefits and limitations. While transfection approaches are capable of generating considerable breakpoint data in a relatively short amount of time, the finite number of constructs used in transfections is a narrow representation of what occurs biologically. To the approach's credit, however, these experiments have been instrumental because studies of naturally occurring NHR events have been limited by the difficulty involved in identifying them. Perhaps unsurprisingly, prior to the availability of large genomic datasets, disease phenotypes led to the identification of many NHR events. As a result, most of our current knowledge regarding naturally occurring examples is disease-related. Though the disease cases are more interesting medically, their investigations provide

pertinent information for protein evolution; each additional reported rearrangement sheds light on the spectrum of possible mutational mechanisms. Fortunately, the large increase in whole genome sequence data is quickly lowering this identification hurdle. Whole genome comparisons, both within and between species, are enabling the identification of numerous rearrangements on which to carry out detailed sequence analyses and further experiments [58–60].

Through the combined efforts of transfection experiments and the study of disease-related NHR events, as well as molecular and biochemical approaches, a collection of motifs and elements that are commonly enriched at or near NHR breakpoints have been identified [30, 52, 56, 57, 61–69]. Several of these are known to be recognized by particular enzymes (Topo I and DNA polymerase, for example [56, 57, 62, 66, 69, 70]) or are thought to lead to DSB-prone structural changes in chromatin [30, 52, 66, 68]. The remaining motifs may also be prone to DSB or may possess enzymatic signals yet to be elucidated. A major goal of these studies is to connect sequence-based data with molecular genetic pathways so that a fuller mechanistic understanding of NHR events can be gained. These identified motifs provide useful sequences to search for when examining rearrangement breakpoints.

A Recent Example: Insights from Hun, a Young Gene Generated by NHR

We recently reported on a young chimeric gene, *Hun,* which was generated by NHR and fixed in the common ancestor of *D. simulans*, *D. mauritiana*, and *D. sechellia* [18]. Its finding provided the unique opportunity to observe the early stages of a new chimeric protein generated by DNA-level recombination. The characterization of its structure, expression, and evolutionary genetics has cast light on the role of nonhomologous recombination in the duplication of a sequence at an ectopic position.

In a large-scale effort to identify new genes, a combination of fluorescent in situ hybridizations (FISH), Southern hybridizations and BLAST techniques were applied across the *D. melanogaster* subgroup. *Hun* was identified as a partial duplicate of the *Bällchen* gene (1867 bp), which arose in the common ancestor of *D. simulans*, *D. sechellia*, and *D. mauritiana*, 2–3 mya. It was shown that the *Hun* duplication involved a t(3R;X) translocation. The total amount of DNA translocated to the X chromosome was ~1520 bp. The derived *Bällchen* coding region was truncated at the 3′ end by ~412 bp, and included only ~65 bp 5′ of the original start codon. Subsequently, ~99 bp were recruited into the 5′ coding region, with an additional ~167 bp to the polyadenylation site. In addition, ~400 bp 5′ of the paralogous *Bällchen* start site were recruited (Fig. 4a). Interestingly, this 5′ UTR contains a putative intron of 49 bp that appears to have evolved de novo.

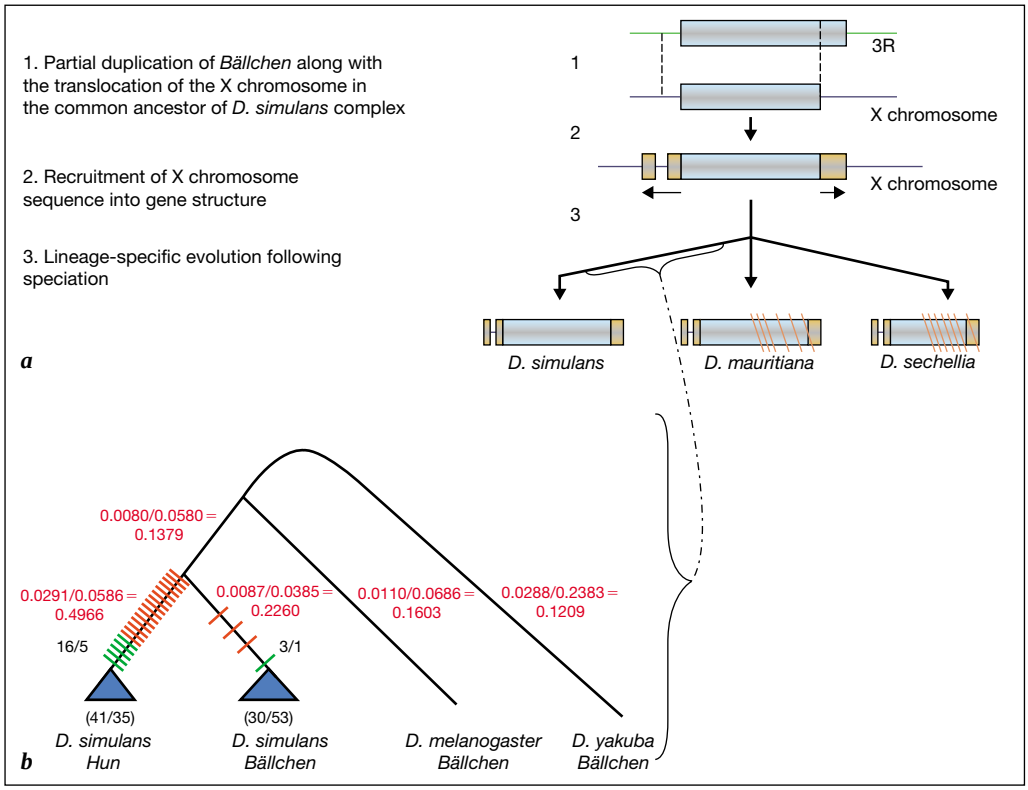


Fig. 4. Models depicting the evolution and population genetics of *Hun*. ***a*** A simplified 3-step model for the origin of *Hun*. Blue bars represent coding regions, brown bars represent UTR regions, and red dashes represent premature stop codons. ***b*** A gene tree for *D. simulans' Hun* and *D. simulans, D. melanogaster,* and *D. yakuba's Bällchen.* The tree displays measurements of divergence as measured by Ka/Ks (red ratios), nonsynonymous and synonymous fixations found along the *Hun* and *Bällchen* branches depicted by colored bars (red represents nonsynonymous changes and green represents synonymous changes, black ratio), and polymorphism found in the *D. simulans* population data (black ratios below triangles, nonsynonymous/synonymous). Taken from [18].

To investigate the mechanism that led to the duplication and translocation of *Hun,* its flanking sequence was examined. In particular, having ruled out the role of an RNA intermediate due to the maintained intron within the coding region, and lack of a poly-A tract, there was interest in the possibility that LCRs may have aided in the rearrangement. No evidence for transposable elements existed near the regions where identity is lost between *Hun* and *Bällchen*. In addition, no evidence for direct repeats was found. When *Bällchen* and

Hun's flanking regions are aligned, only short spurious stretches of identity exist. This lack of evidence for an intermediate RNA step, LCRs, or any other significant sequence identity led to the conclusion that *Hun* originated by an NHR event.

A translocation model proposed by Richardson et al. [71] and later used to explain several rearrangements in a human translocation dataset [58] may provide useful insight for understanding the duplication and translocation of *Hun*. This model accounts for interchromosomal recombination and duplication while avoiding crossovers. In it, recombination occurs between nonhomologous chromosomes through the NAHR of LCRs, such as *Alus* [58]. According to this model, a DSB occurs in one of the two chromosomes (the X chromosome for the *Hun* scenario) near the LCR, followed by strand invasion of homologous sequence belonging to the intact chromosome (chromosome 3R). Strand extension would carry on for some length before rejoining its own chromosome (the X chromosome) at either more distal regions of homology or nonhomology. *Hun's* scenario differs from the previous cases in that we suggest that the initial recombination event between chromosome 3R and the X occurred between regions without any LCRs.

Sequence analyses of *Hun* from *D. simulans*, *D. sechellia,* and *D. mauritiana* revealed that the gene structure has evolved differently in each species [18]. *D. simulans* maintains a single open reading frame, while both *D. sechellia* and *D. mauritiana* have sustained deletions leading to seven and six premature stop codons, respectively. In *D. sechellia,* three significant deletions have occurred in the center of the gene. In *D. mauritiana,* the frame shift was caused by a single base deletion (Fig. 4a). Somewhat surprisingly, screens for the deletions in additional *D. sechellia* and *D. mauritiana* lines suggest that they are fixed. Given the young age of *Hun*, these mutations fixed in a rather short time span.

Along with structural changes, *Hun* has experienced expression evolution. *Bällchen* was shown to be expressed in both sexes in all species within the *D. melanogaster* subgroup. *Hun's* expression, on the other hand, is limited to males in *D. sechellia*, *D. mauritiana*, and *D. simulans*. Tissue-specific RT-PCR revealed that the gene's expression is testes-specific for each of the three species.

Molecular evolution and population genetic analyses were carried out to examine the role that selection has played on *Hun*. Both divergence and polymorphism-based measurements indicated that *Hun* is currently under purifying selection. To infer the role of selection in *Hun's* past, *D. simulans* population data was used. Though standard tests for selection based polymorphism frequency spectrum of *D. simulans* (Fu and Li's D [72], Fu and Li's F [72], and Tajima's D [73]) were nonsignificant alone, the McDonald-Kreitman Test [74] revealed a significant excess of amino acid replacement substitutions along the *Hun* branch (Fig. 4b). These results, combined with the expression

data, were taken as evidence that positive selection for a novel sex-related function drove the fixation of these substitutions.

HR As a Way of Generating Chimeric Proteins – NAHR

So far the evolutionary consequences that can arise through recombination events between chromosomal regions that largely lack sequence identity have been discussed. We now move on to discuss the growing recognition for the role of HR in the formation of chimeric gene structures.

Similar to the approaches used to study NHR, research that has focused on the role that LCRs play in genome rearrangements have primarily been based on naturally occurring disease-related cases [29]. However, with the availability of high quality genomes such as the human genome [75], a more general understanding of the evolutionary role that LCRs play is coming to light [28, 59, 60]. For example, there is growing evidence in primates that LINEs and SINEs are important in mediating rearrangements. Importantly, these rearrangements are not necessarily disease-related but instead have likely produced non-deleterious chimeric proteins as well as making more general evolutionary contributions [58, 60, 76] to primate genome architecture.

A striking example that illustrates the importance of LCRs in producing chimeric gene structures is found in primate *Alu* elements. *Alu* elements are the most numerous members of the SINE family of transposable elements and have been tied to several well-characterized genomic disorders [54, 64, 76]. Until recently, little was known about *Alu*'s more general evolutionary role in shaping genomic architecture. The picture greatly expanded with Bailey et al.'s [59] first fine-scale chromosome-wide analysis of segmental duplications within human chromosome 22. This study resulted in the identification of a surprising number of recent duplication events that resulted in 11 putative chimeric transcripts. Upon further examination of chromosome 22, Babcock et al. [58] reported on numerous *Alu*-related rearrangements including transpositions and duplications, some of which were involved in known chimeric structures. The highly non-random association between *Alus* and rearrangement breakpoints strongly suggested an expansive role for them throughout the genome. Bailey et al. [60] provided strong support for this conjecture through a genome-wide analysis of segmental duplication junctions, 9,464 in total. Out of these duplications, 27% of them had a breakpoint contained within an *Alu*.

Transposable Elements as 'Fragment Joiners'

Current evidence suggests that an important mechanism behind the generation of novel chimeric genes at the DNA level in plants is through the activity of transposable elements. Surveys of the completed genomic sequences of several angiosperms have uncovered a high abundance of MULEs, along with a

subfamily of MULEs that carry gene fragments between terminal inverted repeats. This latter subfamily of MULEs has been named Pack-MULEs [31]. In plant species, Pack-MULEs have been identified in maize [77, 78], rice [79], and *Arabidopsis* [80]. A genome-wide search within rice has identified over 3,000 Pack-MULEs that contain gene fragments averaging 325 bp (with a range of 47–986 bp). Overall, these Pack-MULEs contain DNA fragments from more than 1,000 functional genes. Further, it is estimated that about one-fifth of these identified Pack-MULEs contain DNA fragments from multiple genomic sites and have created novel chimeric gene structures (see fig. 2 for example). At least 5% of these chimeras appear to be functional, with evidence coming from identical full-length cDNAs as well as sequence divergence analyses [31]. In *Arabidopsis*, 5 Pack-MULEs have been identified. The size of the acquired DNA fragments range from 94 to 570 bp and comprise most of the internal DNA of the corresponding elements. However, thus far no Pack-MULE-related chimeric genes in *Arabidopsis* have been identified [26].

Helitrons are a newly identified class of eukaryotic transposable elements that were discovered in the genomes of *A. thaliana*, rice, *Caenorhabditis elegans*, and subsequently in maize [34]. In maize, repeated amplifications and transpositions of *Helitrons* have created numerous gene fragment clusters. Such a system provides intriguing potential for the formation of chimeric gene structures. Illustrating this, Lal et al. [33] recently reported an event in which a *Helitron* had inserted into the maize *Sh2* gene. Though this insertion is believed to have rendered *Sh2* nonfunctional, the example demonstrates a capacity for *Helitron* elements to create chimeric genes.

Chimeric Genes Generated by Gene Fusions

In eukaryotes, two fusion processes have been found to be responsible for the formation of chimeric genes. The first example involved two adjacent genes whereby deletions of the 3′ portion of a 5′ gene and the 5′ portion of a 3′ gene created the novel gene named *Sdic* (S) [36]. The ancestrally adjacent genes were *Cdic* (C) and *AnnX* (A). It was found that the ‘CA gene pair’ tandemly duplicated forming a ‘CACA’ conformation. This was followed by several deletions that eliminated parts of the central ‘AC’ portion, so that a 3′ UTR region belonging to *AnnX* was combined with the 5′ ends of the *Cdic* gene, thus forming a chimeric gene structure. Subsequent evolution within this new gene structure turned intron 3 of *Cdic* into a new promoter region and a new start codon emerged (fig. 3).

The second example is more complex, involving alternative splicing across two adjacent human genes, *KUA* and *UEV* [37]. *KUA* is comprised of 6 exons while *UEV* has 4 exons. Read-through transcription created a large transcript that was alternatively spliced generating the chimerical protein KUA-UEV. The

resultant transcript skipped the sixth *KUA* exon that contained its stop codon, as well as the *UEV* exon that contained its start codon. This case study demonstrates a sophisticated strategy to get rid of stop codons between two tandem genes [16].

Conclusion

Numerous chimeric genes appear to have been generated through DNA level recombination events. While a majority of these genes are ancient, and as result provide limited insight into the mechanism that brought them about, the few young examples that have been collected up to this point demonstrate diverse genetic and evolutionary histories. A major goal for future research is to build upon these young examples and develop a greater understanding of the origination events on a genomic level. With the increasing number of high quality genome projects becoming available, as well as experimental advancements, fundamental estimates for underlying mutational events are becoming feasible both within and between species.

Acknowledgement

This work was supported by a CAS-Max Planck Society Fellowship, a CAS key project grant (No. KSCX2-SW-121), a NSFC award (No. 30325016), a CAS OOCS fund (2004-2-2) and a NSFC key grant (No. 30430400) to W.W.; a USA National Science Foundation CAREER award (MCB0238168) and USA National Institutes of Health R01 grants (R01GM065429-01A1 and 1R01GM078070-01A1) to M.L. at the University of Chicago; a GAANN genomics grant supports J.R.A.

References

1 Kimura M: The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, 1983.
2 Ohno S: Evolution by Gene Duplication. Springer, Berlin, 1970.
3 Betran E, Thornton K, Long M: Retroposed new genes out of the X in *Drosophila*. Genome Res 2002;12:1854–1859.
4 Betran E, Long M: *Dntf-2r*, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. Genetics 2003;164:977–988.
5 Courseaux A, Nahon JL: Birth of two chimeric genes in the Hominidae lineage. Science 2001;291: 1293–1297.
6 Emerson JJ, Kaessmann H, Betran E, Long M: Extensive gene traffic on the mammalian X chromosome. Science 2004;303:537–540.
7 Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H: Emergence of young human genes after a burst of retroposition in primates. PLoS Biol 2005;3:1970–1979.

8 Katju V, Lynch M: On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. Mol Biol Evol 2006;23:11056–11067.
9 Betran E, Emerson JJ, Kaessmann H, Long M: Sex chromosomes and male functions – Where do new genes go? Cell Cycle 2004;3:873–875.
10 Jones C, Custer AW, Begun DJ: Origin and evolution of a chimeric fusion gene in *Drosophila subobscura*, *D. madeirensis* and *D. guanche*. Genetics 2005;170:207–219.
11 Long MY, Langley CH: Natural-selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. Science 1993;260:91–95.
12 Loppin B, Lepetit D, Dorus S, Couble P, Karr TL: Origin and neofunctionalization of a *Drosophila* paternal effect gene essential for zygote viability. Curr Biol 2005;15:87–93.
13 Wang W, Brunet FG, Nevo E, Long M: Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*. Proc Natl Acad Sci USA 2002;99:4448–4453.
14 Wang W, Yu HJ, Long MY: Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. Nat Genet 2004;36:523–527.
15 Brosius J: The contribution of RNAs and retroposition to evolutionary novelties. Genetica 2003;118:99–116.
16 Long M: Evolution of novel genes. Curr Opin Genet Dev 2001;11:673–680.
17 Long M, Betran E, Thornton K, et al: The origin of new genes: Glimpses from the young and old. Nat Rev Genet 2003;4:865–875.
18 Arguello J, Chen Y, Yang S, Wang W, Long L: Origination of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila*. PLoS Genet 2006;2:e77.
19 Gilbert W: Why genes in pieces? Nature 1978;271:44.
20 Roth D, Wilson J: Illegitimate recombination in mammalian cells; in Kucherlapati R, Simth GR (eds): Genetic Recombination. American Society for Microbiology, Washington DC, 1988, pp 621–653.
21 Roth DB, Porter TN, Wilson JH: Mechanisms of nonhomologous recombination in mammalian cells. Mol Cell Biol 1985;5:2599–2607.
22 Roth DB, Wilson JH: Nonhomologous recombination in mammalian cells – role for short sequence homologies in the joining reaction. Mol Cell Biol 1986;6:4295–4304.
23 Allen C, Halbrook J, Nickoloff JA: Interactive competition between homologous recombination and non-homologous end joining. Mol Cancer Res 2003;1:913–920.
24 Roth D, Wilson JH: Relative rates of homologous and nonhomologous recombination in transfected DNA. Proc Natl Acad Sci USA 1985;82:3355–3359.
25 Schwartz M, Zlotorynski E, Goldberg M, Ozeri E, Rahat A, et al: Homologous recombination and nonhomologous end-joining repair pathways regulate fragile site stability. Genes Dev 2005;19:2715–2726.
26 Yu V, Koehler M, Steinlein C, Schmid M, Hanakahi LA, et al: Gross chromosomal rearrangements and genetic exchange between nonhomologous chromosomes following *BRCA2* inactivation. Genes Dev 2000;14:1400–1406.
27 Alexander JRB, Schiestl RH: Homologous recombination as a mechanism for genome rearrangements: environmental and genetic effects. Hum Mol Genet 2000;9:2427–2334.
28 Stankiewicz P, Lupski JR: Molecular-evolutionary mechanisms for genomic disorders. Curr Opin Genet Dev 2002;12:312–319.
29 Shaw C, Lupski JR: Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum Mol Genet 2004;13:R57–R64.
30 Stankiewicz P, Shaw CJ, Dapper JD, Wakui K, Shaffer LG, et al: Genome architecture catalyzes nonrecurrent chromosomal rearrangements. Am J Hum Genet 2003;72:1101–1116.
31 Jiang N, Bao Z, Zhang X, Eddy S, Wessler S: Pack-MULE transposable elements mediate gene evolution in plants. Nature 2004;431:569–573.
32 Bennetzen J: Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 2005;15:621–627.
33 Lal SK, Giroux MJ, Brendel V, Vallejos E, Hannah C: The maize genome contains a Helitron insertion. Plant Cell 2003;15:381–391.
34 Kapitonov V, Jurka J: Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 2001;98:8714–8719.

35 Burns D, Horn V, Paluh J, Yanofsky C: Evolution of the tryptophan synthetase of fungi. Analysis of experimentally fused *Escherichia coli* tryptophan synthetase alpha and beta chains. J Biol Chem 1990;265:2060–2069.
36 Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL: Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature 1998;396:572–575.
37 Thomson T, Lozano JJ, Loukili N, Carrió R, Serras F, et al: Fusion of the human gene for the polyubiquitination coeffector UEV1 with *Kua*, a newly identified gene. Genome Res 2000;10: 1743–1756.
38 Banyai L, Varadi A, Patthy L: Common evolutionary origin of the fibrin-binding structures of fibronectin and tissue-type plasminogen activator. FEBS Lett 1983;163:37–41.
39 Li WH: Molecular Evolution. Sinauer, Sunderland, MA, 1997.
40 Sudhof T, Goldstein JL, Brown MS, Russell DW: The LDL receptor gene: a mosaic of exons shared with different proteins. Science 1985;22:815–822.
41 Patthy L: Modular assembly of genes and the evolution of new functions. Genetica 2003;118: 217–231.
42 Patthy L: Genome evolution and the evolution of exon-shuffling – a review. Gene 1999;238: 103–114.
43 Patthy L: Protein Evolution By Exon-Shuffling. Springer, New York, 1995.
44 Patthy L: Intron-dependent evolution: preferred types of exons and introns. FEBS Lett 1987; 214:1–7.
45 Hedges S: The origin and evolution of model organisms. Nat Rev Genet 2002;3:838–849.
46 Long M, de Souza SJ, Rosenberg C, Gilbert W: Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. Proc Natl Acad Sci USA 1996;93: 7727–7731.
47 Kaessmann H, Zollner S, Nekrutenko A, Li WH: Signatures of domain shuffling in the human genome. Genome Res 2002;12:1642–1650.
48 Liu M, Grigoriev A: Protein domains correlate strongly with exons in multiple eukaryotic genomes–evidence of exon shuffling? Trends Genet 2004;20:399–403.
49 Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD: Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. Trends Genet 2005;21: 326–329.
50 Shendure J, Church GM: Computational discovery of sense-antisense transcription in the human and mouse genomes. Genome Biol 2002;3:research0044.1–0044.14.
51 Fu Y, Yu JC, Cheng TC, Lou MA, Hsu GC, et al: Breast cancer risk associated with genotypic polymorphism of the nonhomologous end-joining genes: a multigenic study on cancer susceptibility. Cancer Res 2003;63:2440–2446.
52 Nobile C, Toffolatti L, Rizzi F, Simionati B, Nigro V, et al: Analysis of 22 deletion breakpoints in dystrophin intron 49. Hum Genet 2002;110:418–421.
53 Zucman-Rossi J, Legoix P, Victor JM, Lopez B, Thomas G: Chromosome translocations based on illegitimate recombination in human tumors. Proc Natl Acad Sci USA 1998;95:11786–11791.
54 Hu X, Worton RG: Partial gene duplication as a cause of human disease. Hum Mutat 1992;1:3–12.
55 Allgood ND, Silhavy TJ: Illegitimate recombination in bacteria; in Kucherlapati R, Simth GR (eds): Genetic Recombination. American Society for Microbiology, Washington, DC, 1988.
56 van Rijk A, Bloemendal H: Molecular mechanisms of exon shuffling: illegitimate recombination. Genetica 2003;118:245–249.
57 van Rijk AAF, de Jong WW, Bloemendal H: Exon shuffling mimicked in cell culture. Proc Natl Acad Sci USA 1999;96:8074–8079.
58 Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, et al: Shuffling of genes within low-copy repeats on 22qll (LCR22) by Alu-mediated recombination events during evolution. Genome Res 2003;13:2519–2532.
59 Bailey J, Yavor AM, Viggiano L, Misceo D, Horvath JE, et al: Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. Am J Hum Genet 2002;70: 83–100.
60 Bailey J, Liu G, Eichler EE: An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet 2003;73:823–834.

61 Abeysinghe S, Chuzhanova N, Krawczak M, Ball EV, Cooper DN: Translocation and gross deletion breakpoints in human inherited disease and cancer I: nucleotide composition and recombination-associated motifs. Hum Mutat 2003;22:229–244.
62 Been M, Burgess RR, Champoux JJ: Nucleotide sequence preference at rat liver and wheat germ type 1 DNA topoisomerase breakage sites in duplex SV40 DNA. Nucleic Acids Res 1984;12: 3097–3114.
63 Borgato L, Bonizzato A, Lunardi C, Dusi S, Andrioli G, et al: A 1.1-kb duplication in the p67-*phox* gene causes chronic granulomatous disease. Hum Genet 2001;108:504–510.
64 Chi C, Tsai CR, Chen LH, Lee HF, Mak BSC, et al: Maple syrup urine disease in the Austronesian aboriginal tribe Paiwan of Taiwan: a novel DBT (E2) gene 4.7 kb founder deletion caused by a nonhomologous recombination between LINE-1 and Alu and the carrier-frequency determination. Eur J Hum Genet 2003;11:931–936.
65 Chou C, Morrison SL: A common sequence motif near nonhomologous recombination breakpoints involving Ig sequences. J Immunol 1993;12:5350–5360.
66 Kumatori A, Faizunnessa NN, Suzuki S, Moriuchi T, Kurozumi H, Nakamura M: Nonhomologous recombination between the cytochrome b_{558} heavy chain gene (*CYBB*) and LINE-1 causes an X-linked chronic granulomatous disease. Genomics 1998;53:123–128.
67 Nikiforov Y, Koshoffer A, Nikiforova M, Stringer J, Fagin JA: Chromosomal breakpoint positions suggest a direct role for radiation in inducing illegitimate recombination between the *ELE1* and *RET* genes in radiation-induced thyroid carcinomas. Oncogene 1999;18:6330–6334.
68 Singh G, Kramer JA, Krawetz SA: Mathematical model to predict regions of chromatin attachment to the nuclear matrix. Nucleic Acids Res 1997;25:1419–1425.
69 Zhu J, Schiestl RH: Topoisomerase I involvement in illegitimate recombination in *Saccharomyces cerevisiae*. Mol Cell Biol 1996;16:1805–1812.
70 Tanizawa A, Hohn KW, Pommier Y: Induction of cleavage in topoisomerase I c-DNA by topoisomerase I enzyme from calf thymus and wheat germ in the presence and absence of camptothecin. Nucleic Acids Res 1993;21:5157–5166.
71 Richardson C, Moynahan ME, Jasin M: Double-strand break repair by interchromosomal recombination: suppression of chromosomal translocations. Genes Dev 1998;12:3831–3842.
72 Fu YX, Li WH: Statistical tests of neutrality of mutations. Genetics 1993;133:693–709.
73 Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989;123:585–595.
74 McDonald JH, Kreitman M: Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 1991;351:652–654.
75 Venter J, Adams MD, Myers EW, Li PW, Mural RJ, et al: The sequence of the human genome. Science 2001;291:1304–1351.
76 Batzer M, Deininger PL: Alu repeats and human genomic diversity. Nat Rev Genet 2002;3: 370–379.
77 Bennetzen J, Springer PS: The generation of mutator transposable element subfamilies in maize. Theor Appl Genet 1994;87:657–667.
78 Talbert L, Chandler VL: Characterization of a highly conserved sequence related to mutator transposable elements in maize. Mol Biol Evol 1988,5:519–529.
79 Turcotte K, Srinivasan S, Bureau T: Survey of transposable elements from rice genomic sequences. Plant J 2001;25:169–179.
80 Yu Z, Wright SI, Bureau TE: Mutator-like elements in *Arabidopsis thaliana*. Structure, diversity and evolution. Genetics 2000;156:2019–2031.

Manyuan Long
Department of Ecology and Evolution, University of Chicago
1101 East 57th St., Zoology
301E, Chicago, IL 60637 (USA)
Tel. +1 773 702 0557, Fax +1 773 702 9740, E-Mail mlong@uchicago.edu

Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 147–162

Exaptation of Protein Coding Sequences from Transposable Elements

N.J. Bowen, I.K. Jordan

School of Biology, Georgia Institute of Technology, Atlanta, Ga., USA

Abstract

The activity of transposable elements (TEs) has had a profound impact on the evolution of eukaryotic genomes. Once thought to be purely selfish genomic entities, TEs are now recognized to occupy a continuum of relationships, ranging from parasitic to mutualistic, with their host genomes. One of the many ways that TEs contribute to the function and evolution of the genomes in which they reside is through the donation of host protein coding sequences (CDSs). In this chapter, we will describe several notable examples of eukaryotic host CDSs that are derived from TEs. Despite the existence of a number of such well-established cases, the overall extent and significance of this phenomenon remains a matter of controversy. Genome-scale computational analyses have yielded vastly different estimates for the fraction of host CDSs that are derived from TEs. We explain how these seemingly contradictory findings are the result of specific ascertainment biases introduced by the different methods used to detect TE-related sequences. In light of this problem, we propose a comprehensive and systematic framework for definitively characterizing the contribution of TEs to eukaryotic CDSs.

Copyright © 2007 S. Karger AG, Basel

Transposable Elements Defined

Transposable elements (TEs) are DNA sequences that can move (transpose) from one chromosomal location to another within the genome. Along with the capacity to move around the genome, TEs can replicate themselves and accumulate over time. The transpositional activity of TEs has had a substantial effect on the structure and function of eukaryotic genomes. For instance, TEs can cause phenotypically relevant mutations by inserting in or around genes, and ectopic recombination, mediated by homologous element sequences dispersed throughout the genome, can lead to large scale chromosomal re-arrangements. In addition, TEs are both abundant and ubiquitous. Remnants of TE insertions

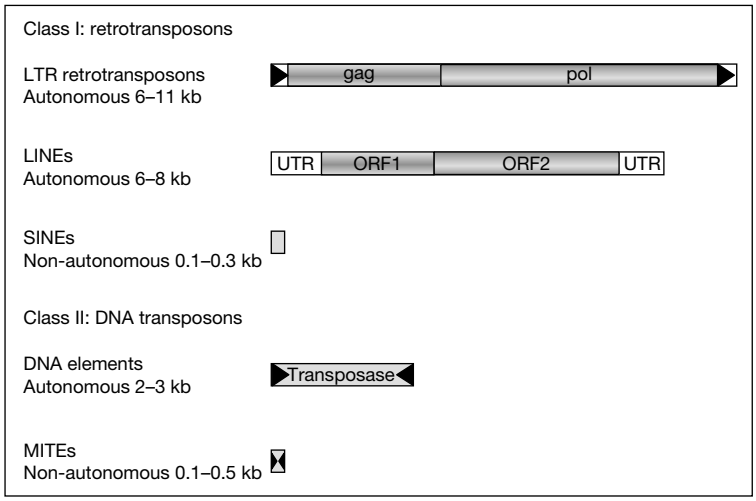


Fig. 1. Classification and structure of TEs. Long terminal repeat (LTR) retrotransposons are flanked by two direct repeat sequences. They typically possess the gag and pol ORFs, which encode structural (gag) and enzymatic (pol) proteins involved in reverse transcription. LINE elements have untranslated regions (UTRs) of variable length on either side of two ORFs. ORF1 often encodes a nucleic acid binding protein, while ORF2 encodes reverse transcriptase. SINEs are much shorter non-autonomous retroelements that lack coding capacity. Autonomous DNA-type elements contain two terminal inverted repeats (TIRs) surrounding a single open reading frame that encodes the transposase enzyme. Non-autonomous DNA elements, such as MITEs, possess TIRs as well but lack coding capacity.

often make up more than half of any given eukaryotic genome sequence, and TEs have been found in all three domains of life for virtually every organism with characterized genomic sequences. In this chapter, we will focus exclusively on eukaryotic TEs, with an emphasis on human elements that have contributed to the evolution of protein coding sequences (CDSs).

Eukaryotic TEs are categorized into two broad classes [1] (fig. 1). Class I elements, or retroelements, transpose via the reverse transcription of an RNA intermediate. Retroelements include long- and short-interspersed nuclear elements, known as LINEs and SINEs respectively, as well as long terminal repeat (LTR) containing retrotransposons. LINEs encode all the enzymatic machinery necessary for their retrotransposition, while SINEs are non-autonomous elements that lack coding capacity. SINEs are thought to be retrotransposed in trans using enzymes encoded by LINEs [2]. LINEs are the single most abundant class of elements in the human genome making up more than 20% of the sequence [3]. A single family of LINE elements alone, LINE1, has amplified to

more than half a million copies in the human genome. In addition to catalyzing their own transcription, as well as that of SINEs, LINE encoded reverse transcriptase enzymes are probably responsible for generating the majority of processed pseudogenes in the human genome [4].

SINEs originate from non protein-coding RNAs transcribed by RNA polymerase III, tRNAs for the most part, that have been amplified by reverse transcription [5]. SINEs are actually the most numerous class of elements in the human genome, with more than 1.5 million copies identified, but make up a slightly lower overall fraction of the human genome (~13%) than LINEs due to their smaller size [3]. Most of the human genome SINEs are Alu elements. Alus are relatively young elements found exclusively within the primate evolutionary lineage [6]; they originated from 7SL RNA transcripts, which make up part of the ribosomal signal recognition particle [7]. Alus are particularly interesting because, unlike most other human TEs, they have accumulated in relatively GC-rich sequences found in close proximity to genes. An analysis of the age distribution of Alus revealed that this phenomenon is not due to any insertion site preference [3]. Rather it appears that Alus have been preferentially retained at gene encoding loci. This has been taken to suggest that Alus are genomic symbionts that play some beneficial role for the genomes in which they reside [3].

LTR retrotransposons are closely related to retroviruses [8]. They have open reading frames (ORFs) that encode capsid-like proteins, as well as the enzymes involved in retrotransposition, but lack the envelope ORF that confers intercellular infectivity to retroviruses [9]. In fact, LTR retrotransposons in humans are referred to as endogenous retroviruses, and many of them probably evolved from retroviruses that infected primate germline sequences and subsequently lost their infectivity [10]. Most of the LTR retrotransposon sequences in the human genome are found as solo LTRs, which are the result of intraelement LTR-LTR recombination events that excise the internal element sequences. LTR elements make up just under 10% of the human genome [3].

Class II, or DNA-type, elements transpose from DNA-to-DNA via a cut-and-paste mechanism catalyzed by the enzyme transposase. These elements generally contain inverted terminal repeats (TIRs) recognized by the DNA-binding domain of transposase and a single ORF encoding the transposase. Non-autonomous DNA elements containing only TIRs may be transposed in trans by related full length autonomous elements. Miniature inverted-repeat elements (MITEs) are a group of small high copy number non-autonomous DNA-type elements originally discovered in plants [11] and subsequently found in a wide variety of eukaryotic genomes [12]. Like Alus, MITEs are often found in close association with gene sequences and are thus thought to play some beneficial role related to gene regulation. DNA-type elements are the most common class of bacterial transposons, where they are known as IS

elements. DNA-type elements are also particularly abundant among insect and plant genomes, but less so in the human genome where they make up ~3% of the sequence. Despite their relatively low numbers in the human genome, DNA-type elements make up most of the known cases of TE-derived human CDSs [3]. The reason for this over-representation is currently unknown. It might be due to the fact that the transposase ORF provides a ready-made protein with DNA-binding properties that are particularly useful for the host [13]. For instance, domesticated transposase ORFs could play a role in mitigating the harmful effects of TEs by repressing transposition and/or they could influence the expression of host genes by acting as novel transcription factors.

There are also scattered examples of anomalous TEs that do not fit neatly into either of the two aforementioned broad classes I or II. For instance, DIRS1-like elements encode reverse transcriptase enzymes that are similar to those of LTR retrotransposons, but they lack integrase coding capacity as well as LTRs [14]. An even more unusual class of elements found in insects and plants possesses similarities to both non-LTR and LTR retrotransposons [15]. Some of these so-called Penelope-like elements do have LTRs, but they may be inverted in orientation. Many are 5′ truncated like non-LTR elements, and the reverse transcriptases of these elements are interrupted by an intron and most similar to the enzyme telomerase. The increasing appearance of such difficult to classify elements suggests the need for a revised classification scheme for TEs, an issue which has been addressed recently [1].

Selfish DNA Theory of TEs

The recognition of TEs' broad distribution and high copy numbers, i.e. their evolutionary success, in eukaryotic genomes posed an explanatory challenge to biologists. Many wondered which attributes of the elements could best explain their ubiquity. This line of inquiry was based on a classic neo-Darwinian mode of thought, which held that the success of a gene must be predicated upon the utility that it provided for the organisms that encoded it. If this paradigm held for TEs, then it follows that they must be playing important and demonstrable roles for the genomes in which they reside. Thus, the first impulse for investigators interested in explaining the presence of TEs was to posit potential adaptive benefits that they may provide to their host organisms. In 1980, two seminal papers, published back-to-back in Nature [16, 17], completely inverted this explanatory paradigm for the presence and success of TEs.

These two papers laid the foundation for what is known as the selfish DNA theory of TEs. The selfish DNA theory holds that TEs are essentially genomic parasites, which serve only to increase their own abundance even at the expense

of their host genomes. The authors pointed out that existence of TEs can be explained solely by virtue of their ability to out-replicate their host genomes. This is because, in addition to being transmitted vertically like standard host genes, TEs are also replicated within the genome when they transpose. This replicative component of their life cycle gives TEs an inherent fitness advantage relative to host genes transmitted in a strict Mendelian fashion. This replicative advantage alone, with no regard whatsoever to any functional role or adaptive benefit that they may provide to their hosts, is sufficient to explain their evolutionary success. Later it was shown that TEs can spread within genomes and populations even in the face of deleterious effects that they may exert, via insertional mutations for example, on their host genomes [18]. This finding further emphasized the potentially parasitic nature of TEs.

The selfish DNA theory of TEs represented a true paradigm shift and continues to play an important role as a null hypothesis for the understanding of TEs' evolutionary significance. This is important in the sense that it helps to avoid the kind of tautological adaptationist thinking whereby the mere presence of a biological feature demands a plausible adaptive explanation. On the other hand, this new paradigm for TEs, while logically unassailable, also had a chilling effect on investigations into any functional role, or adaptive benefit, that TEs may play for their host genomes. In retrospect, it is interesting to note the divergent tacks taken by the authors of the two selfish DNA papers with regard to the potential functional roles played by TEs. One group applied a more cautious and measured approach being careful to point out that the selfish nature of TEs would not necessarily preclude them from being co-opted to play some beneficial role for their hosts [17]. However, the second group advocated a much harder line pointing out that the selfish nature of TEs rendered the consideration of any functional role that they may play ultimately futile [16]. Of course these two aspects of TE biology – selfish versus adaptive – are not mutually exclusive, and in recent years a more balanced perspective, which holds that TEs exist on a continuum from strict parasitism to mutualism, has emerged [19].

Molecular Domestication

While the selfish DNA theory can still be considered as the null hypothesis by which the presence of TEs is explained, there are by now many exceptions to this view on TEs' evolutionary and functional significance, or lack thereof [19]. Wolfgang Miller coined the phrase 'molecular domestication' to describe the process whereby a formerly selfish TE is co-opted to perform a function that benefits its host genome [20]. Molecular domestication of TEs is

an example of the more general phenomenon of exaptation. The term exaptation was introduced by Gould and Vrba to account for a biological adaptation that plays a current role distinct from the original function that was selected for [21]. In the case of TEs, exaptations result from selection pressures exerted at different levels of biological organization. On the one hand, TE CDSs originally evolve under selection pressure at the genomic level. While on the other, the selection that governs the evolution of host gene sequences is exerted at the organismic level. In both cases, the selection is based on differential reproductive success. Organismic level selection is based on differential reproductive success of individuals in a population. In order to get established within a genome, TEs face selection pressure to transpose, i.e. reproduce in the genome, efficaciously and thus out-reproduce both their host genomes and other competing elements. However, this evolutionary mode does not represent an effective long-term strategy since transposition is a highly mutagenic process that often causes deleterious changes to the host genome. Unchecked transposition and accumulation of element sequences could ultimately lead to the extinction of the TEs' host evolutionary lineage, which in turn would mean extinction of the elements themselves. To counter this possibility, TEs have evolved a number of mechanisms that mitigate the harmful effects of transposition. For instance some TEs, such as LINEs in human and mouse, confine their expression to germline tissues [22]. This helps to ensure transmission of newly replicated elements to future generations while minimizing the harmful effects of somatic mutations caused by transposition. As another example of host-element co-evolution, P-elements in *Drosophila* have evolved a strategy of self-regulation by encoding a repressor protein that blocks transposition in already infected genomes [23]. Of course, the ultimate co-evolutionary strategy exhibited by TEs is molecular domestication whereby the formerly selfish element sequences make themselves indispensable to their host genome by taking on some essential functional role.

There is a growing body of evidence that demonstrates a number of different ways that TEs have evolved from strictly parasitic elements to mutualistic sequences that benefit their host genomes. For instance, there are numerous documented cases where TEs have been shown to donate regulatory sequences that control the expression of nearby host genes [24–26]. For the rest of this chapter though, we will focus exclusively on the cases where formerly selfish TE sequences have been domesticated to provide CDSs for their eukaryotic host genomes. The extent and overall significance of this phenomenon is currently a matter of some debate. As genome sequences accumulate, more and more examples of TE-derived host CDSs are posited. However, some of these cases have proven illusory and different methods of detection of TE-derived CDSs often yield vastly different results. In addition to providing a few canonical

examples of TE-derived host CDSs, we will explore issues pertaining to ascertainment biases at play in their discovery.

Host CDSs from TEs

In this section, we will briefly outline several canonical examples of how TE ORFs have been exapted as host gene CDSs – Telomerase, RAG recombinase and SETMAR. Telomerase is an enzyme that helps to replace DNA residues that are inevitably lost from the ends of eukaryotic chromosomes during their replication [27]. Telomerase uses RNA oligonucleotides as templates for the addition of DNA sequences to chromosome ends; in other words, it is a reverse transcriptase (RT). Sequence analysis of telomerase revealed that it shares a common evolutionary origin with the RTs of retrotransposable elements [28]. Phylogenetic comparison of telomerase with the RT domains of TEs indicates that the telomerase RT probably diverged from LINE-like elements in early eukaryotic evolution, which points to the exaptation of this critical cellular activity from a class of formerly selfish elements [28].

The RAG recombinase enzymes – RAG1 and RAG2 – are together responsible for catalyzing V(D)J recombination in vertebrate genomes [29]. V(D)J recombination is the mechanism by which vertebrates generate immunological diversity in antibody and T-cell receptor molecules [30]. The breaking and rejoining of chromosomal segments catalyzed by the RAG enzymes allows for the production of a vastly diverse array of immunoglobulin encoding sequences, which is capable of countering the diversity of pathogens that challenge the immune system. The striking similarity between the processes of RAG catalyzed V(D)J recombination and transposition of DNA-type elements led to the suggestion of an evolutionary relationship between the two [31]. This proposition was later confirmed by experimental work showing that the RAG recombinases can catalyze transposition within and between chromosomes [32]. Recently, a direct evolutionary link between the RAG1 sequence and a family of DNA-type elements has been established [33]. Thus, it appears likely that the ability of the vertebrate immune system to generate immunogenic diversity evolved from TEs as well. The implications of this particular exaptation event for the survival of the vertebrate evolutionary lineage are striking.

The human SETMAR protein provides a more recent example of the exaptation host CDSs from a TE [34, 35]. Because this particular domestication event took place in the more recent evolutionary past, investigators were able to more definitively characterize the relationship between the TE and its related host gene. Indeed, SETMAR was originally characterized as a chimeric transcript that combined a SET methyltransferase domain with the transposase domain

from a specific family of DNA-type element named Hsmar1 [35]. Comparative analysis of corresponding genomic regions cloned from related primates revealed that this particular domestication event occurred between 40–58 million years ago after the Hsmar1 element inserted downstream from the SET domain encoding exons [34]. The function of this particular domesticated gene remains less well understood than is the case for the more ancient exaptation events that led to evolution of telomerase and RAG recombinase from TE sequences. Investigators were able to demonstrate that the catalytic activity of the transposase derived domain of SETMAR has been lost while the DNA binding activity remains [34]. This raised the intriguing possibility that the evolution of SETMAR may have facilitated the de novo emergence of complex regulatory network involving the binding of SETMAR to numerous Hsmar1 derived TIR binding sites dispersed throughout the genome.

The three selected examples of TE-CDS domestication described here represent only a fraction of the known cases. Another noteworthy example is the case of Daysleeper, a DNA-type element that has been domesticated in *Arabidopsis* and shown to be essential for plant development by virtue of the regulatory effects that it exerts on numerous genes [36]. In humans, both the centromere binding protein gene (*CENBP*) and the Jerky gene are derived from related DNA-like elements [37–39]. An exhaustive description of all such cases is outside the scope of this manuscript. However, several other reports do provide a more in depth accounting of protein coding sequences exapted from TEs [3, 37, 39, 40].

Genome Wide Analyses

Despite these solid examples of TE contributions to CDSs, the extent of this phenomenon remains a matter of substantial controversy. With the accumulation of eukaryotic genome sequences, a number of attempts have been made to exhaustively characterize instances of TE-derived host CDSs [3, 37, 39, 41, 42]. One of the open questions that such studies address is the proportion of host CDSs that can be demonstrated to have evolved from TEs. Prior to the completion of the human genome sequence there were 20 known cases of human CDSs derived from TEs [37, 39]. Analysis of the draft sequence of the human genome found 27 additional cases yielding a total of 47 distinct human TE-derived CDSs [3]. This figure represents a fairly negligible fraction (~0.16%) of all human genes, given the lower bound estimate for the human gene count (30,000) reported at that time [3]. The same year however, using similar detection techniques on a set of human gene sequences from a different source, Nekrutenko and Li published their own genome-scale analysis where they reported that ~4%

of analyzed human genes had CDSs that were derived from TEs [42]. Clearly such vastly different estimates call for some sort of reconciliation.

Pavlicek and colleagues took another look at the findings of Nekrutenko and Li and pointed out several caveats that should be taken into consideration when trying to determine the extent of CDSs derived from TEs [43]. First of all, they found that, of the set of genes identified by Nekrutenko and Li as TE-derived, 30% were annotated as hypothetical and 63% were annotated as predicted. In other words, there was no experimental evidence that supported these particular genes as being bona fide functional CDSs. In addition, the majority of human CDSs that Nekrutenko and Li found with similarity to TEs were derived from Alu (SINE) elements that lack protein coding capacity. Pavlicek et al. found this suspicious as well, since the vast majority of CDSs previously reported to contain Alu related sequences have only been detected at the mRNA level. In light of these issues, Pavlicek et al. took a far more conservative approach to identifying TE-derived CDSs; specifically, they analyzed CDSs taken only from representative 3D structures. The 3D structures represent the most accurate source evidence for the actual existence of the proteins under consideration. When these CDSs were analyzed using the same detection technique as Nekrutenko and Li no evidence of TE-derived sequences was found. A slightly more sensitive technique did reveal 28 cases of TE-derived CDSs but all of these came from TEs that are known to encode proteins and none were from Alu elements, which lack protein coding capacity.

Despite the apparent absence of SINE related sequences among CDSs with representative 3D structures, the facts remain that Alu elements are both highly prevalent in human gene-rich regions [3] and harbor numerous potential splice sites that can facilitate their incorporation into mRNA transcripts [44]. Thus, Alus would appear to be ideally positioned to be integrated into the CDSs of host genes. Indeed, numerous alternatively spliced human exons (~5%) were found to contain Alu sequences [44]. Several individual cases of how Alu sequences have become 'exonized' have been explicated in detail, revealing the evolutionary timing of these events as well as the specific mutations that led to their incorporation into gene transcripts [45]. However, the actual protein coding potential and biological function of these sequences is still an open question. Comparative sequence analysis, establishing both the conservation of Alu-derived open reading frames and a conservative pattern of sequence substitution [46] should help to settle this matter.

Consistent with the conservative approach of Pavlicek et al., a more recent publication from the Nekrutenko group refuted one of their own earlier discoveries of a mouse CDS that appeared to be derived almost entirely from SINEs [47]. Comparative sequence analysis with other *Mus* species, as well as the rat, did not find any evidence for the conservation of the ORF of this TE-derived

gene. The authors concluded that the original discovery represented an artifact and emphasized the importance of validation of computationally predicted TE-derived CDSs. This report also underscored the paucity of examples of non-artifactual validated cases where non-coding TEs, such as SINEs, contribute CDSs to host genes [48]. Rather, it appears that the vast majority of well-supported cases of TE-derived host CDSs come from TEs that already encode proteins.

A New Framework is Needed

The contradictory results outlined in the previous section illustrate the confusion and controversy that still surround the issue of TE-derived CDSs [49]. We would like to close this chapter by arguing that a new framework, one that is both comprehensive and systematic, is needed to understand the extent and biological significance of TE-derived CDSs. It is worth noting here that the extent of TE-derived CDSs may not necessarily be directly related to its evolutionary significance. For instance, even if the contribution of TEs to host CDSs turns out to be low in terms of overall numbers, its impact in terms of biology and evolution may still be substantial. The single case of the RAG-recombinases and vertebrate specific immunity underscores this point. It is also tempting to speculate as to whether differing extents of domestication between evolutionary lineages may be responsible for driving increases in complexity that mark the eukaryotic crown group. In any case, a rigorous elucidation of the extent of host protein coding sequences that are derived from TEs will be critical for our understanding of eukaryotic genome structure, function and evolution.

It has occurred to us and others that a substantial problem lies in the differences in sensitivity of the methods used to detect TE-related sequences among protein coding genes. Most studies rely on the widely used RepeatMasker program [50] to detect TE related sequences. RepeatMasker works at the DNA-level and compares genomic sequences to a library of consensus sequences that represent previously characterized TE families. This approach has two distinct disadvantages. First of all it can only detect TEs that are already known. This is not a big problem for well-characterized species such as human, but it may represent a substantial limitation for less well characterized evolutionary lineages. Perhaps even more importantly, RepeatMasker can only detect TE sequences that have diverged relatively recently from other members of the same family. This is partly due to the use of consensus sequences but more so to the reliance on DNA-DNA sequence comparisons. With only four different residues to compare, substitutions between related DNA sequences quickly become saturated and their evolutionary relationships are obscured. Protein sequences, on the other hand, retain the signal of common ancestry for much longer periods of time.

The ascertainment bias that results from the reliance on DNA-DNA sequence comparisons was driven home by a recent re-analysis of TE-derived sequences among human protein coding genes [41]. Roy Britten took the same TE consensus sequence libraries used for RepeatMasker and translated them in all reading frames. These conceptual translations were then used as queries in protein-protein BLAST searches against all human proteins. The protein sequence comparison resulted in a more than two-fold increase from 814 detected TE-derived CDSs to 1,950 such cases. These newly detected cases represent more ancient associations between TE-derived sequences and human genes and in that sense may be even more likely to be validated with experimental and/or 3D structural information. In addition, while protein-protein sequence comparisons are more sensitive than DNA-DNA comparisons, there are even more sensitive ways to search for common ancestry between sequences such as profile comparisons, using position specific score matrices (PSSMs) or hidden Markov models (HMMs), and direct comparisons between 3D protein structures, which are the most sensitive methods of all. The use of such techniques would undoubtedly turn up additional cases of TE-derived, or at the very least TE-related, host CDSs.

A favorite example of ours can serve to further illuminate the challenges for uncovering relationships between TEs and host CDSs. PAX8 is a nuclear transcription factor that is involved in thyroid and kidney development and implicated in the etiology of several different cancers. PAX8 is a member of the paired box (PAX) family of transcription factors that are expressed in cell specific patterns during metazoan development [51, 52]. PAX proteins contain an amino-terminal, sequence specific DNA binding domain known as the paired box, which consists of tandem helix-turn-helix (HTH) motifs [53, 54]. Protein sequence-based homology searches have uncovered a highly significant similarity between the paired box domain and the transposase present in members of the Tc1 family of TEs [55, 56]. Structural modeling has likewise shown the presence of two HTH motifs in the Tc1 transposase sequences [55, 57]. The similarity between Tc1 transposase and the paired box domain is so reliable that transposase sequences are now routinely used as an outgroup to root phylogenetic comparisons of within and between species comparisons of PAX proteins [55, 58].

There are 9 PAX genes in the human genome and many more genes that encode domains with HTH motifs that may be distantly related to transposase domains. However, the PAX genes in particular are present only in the animal lineage of eukaryotes; they have not been found in unicellular eukaryotes, fungi, plants nor in prokaryotes [59]. This lineage-specificity of PAX genes stands in stark contrast to the widespread distribution of the Class II family of DNA-type elements to which Tc1 elements belong. Based on these disparate

phyletic distributions, a transposase origin of PAX genes is the most likely evolutionary scenario that explains their sequence similarity. Despite their robust and well-characterized relationship, when the PAX8 DNA CDS is run through RepeatMasker no evidence of a TE origin is uncovered. Clearly, a strict DNA-centric genome-scale approach to uncovering TE-derived CDSs will only tell part of the story.

Finally, we would like to emphasize here that the challenge of ascertainment biases inherent to the different methods also presents an important opportunity. Once the particular strengths and weaknesses of different approaches are recognized and considered, a more systematic approach to the detection of TE-derived CDSs can be devised. Specifically, we would like to propose that any genome-scale computationally based attempt to uncover TE-derived host CDSs must involve the use of numerous complementary approaches, each of which is appropriate to its own level of evolutionary relatedness between the TEs and CDSs (fig. 2): (i) DNA-DNA sequence comparison methods can be used to detect recent putative exaptation events followed successively by (ii) protein-protein sequence comparisons, (iii) profile-protein comparisons and (iv) structure-protein or structure-structure comparisons, each of which in turn may reveal more ancient associations between TEs and CDSs. Detection of such relationships must only represent the first step in the process though. Further confidence in the validity of TE-gene associations can be achieved by comparative sequence analyses aimed at detecting both conservation of TE-derived ORFs as well as conservative DNA substitution patterns that are indicative of purifying selection based on protein function. Finally, the kind of phyletic distribution comparison described earlier for the case of PAX8 can be used to definitively establish the evolutionary directionality of the relationship between the TE and host CDSs. The donor sequence set (the transposase in the case of PAX encoding genes) should be characterized by a broader distribution among more distantly related species than the acceptor group of sequences. The breadth of sequence distribution can also be used to inform the direction of the relationship between TE and host gene. In the case of the telomerase for example, its RT represents only a fraction of the sequence diversity of all retrotransposon RTs, which is consistent with its origin from one particular lineage along the RT phylogenetic tree. While seemingly exhaustive, this kind of comprehensive approach that we propose is ideally suited to the computational approach. In fact a recently published paper from the group of Peer Bork proposed an analogous, if more narrow in scope, algorithmic approach aimed at discovering and then validating cases of host CDSs that may be derived from TEs [60]. Hopefully, the problem of the extent and significance of TE-derived host CDSs will yield to such a systematic approach, and in so doing, help us to better understand the ancient and ongoing evolutionary dynamic between TEs and their host genomes.

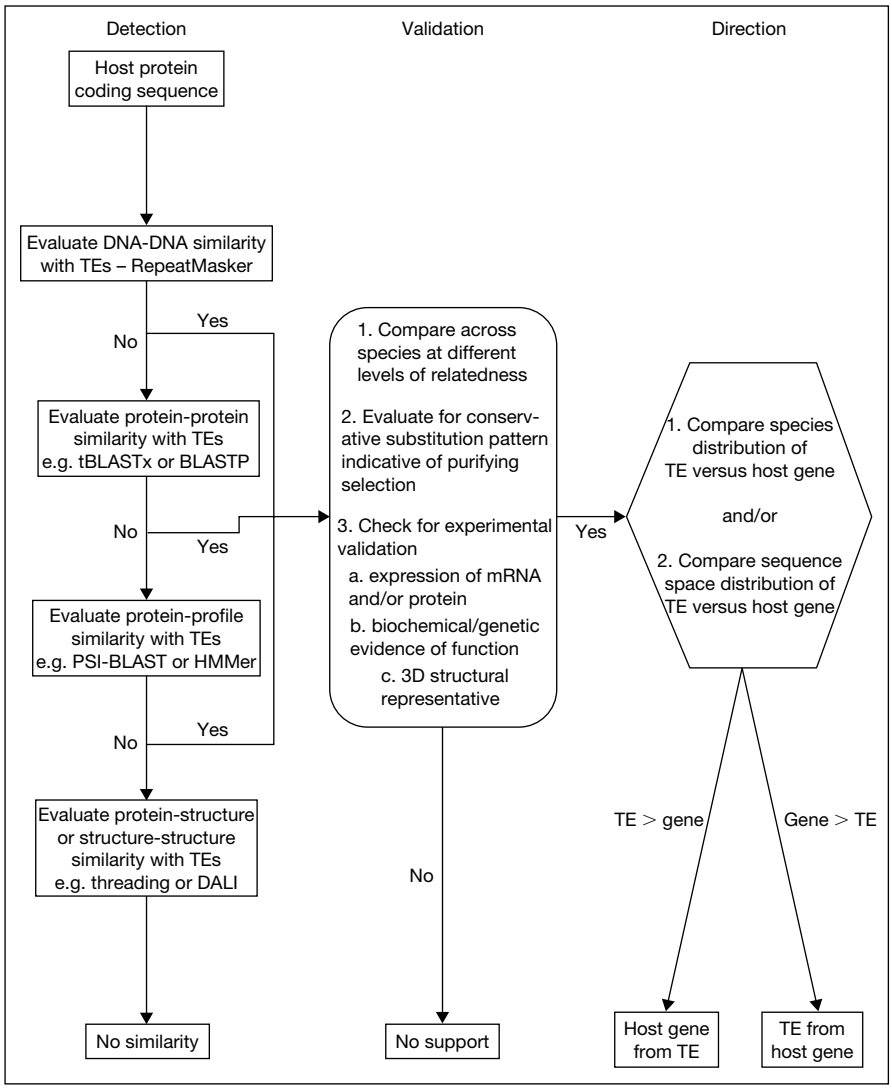

Detection
Validation
Direction
Host protein coding sequence
Evaluate DNA-DNA similarity with TEs – RepeatMasker
No
Yes
Evaluate protein-protein similarity with TEs e.g. tBLASTx or BLASTP
No
Yes
Evaluate protein-profile similarity with TEs e.g. PSI-BLAST or HMMer
No
Yes
Evaluate protein-structure or structure-structure similarity with TEs e.g. threading or DALI
No similarity
1. Compare across species at different levels of relatedness
2. Evaluate for conservative substitution pattern indicative of purifying selection
3. Check for experimental validation
a. expression of mRNA and/or protein
b. biochemical/genetic evidence of function
c. 3D structural representative
No
No support
Yes
1. Compare species distribution of TE versus host gene
and/or
2. Compare sequence space distribution of TE versus host gene
TE > gene
Gene > TE
Host gene from TE
TE from host gene

Fig. 2. Scheme for the detection, validation and characterization of TE-derived host CDSs.

References

1 Capy P: Classification and nomenclature of retrotransposable elements. Cytogenet Genome Res 2005;110:457–461.

2 Dewannieux M, Esnault C, Heidmann T: LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 2003;35:41–48.

3 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al: Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
4 Esnault C, Maestre J, Heidmann T: Human LINE retrotransposons generate processed pseudogenes. Nat Genet 2000;24:363–367.
5 Daniels GR, Deininger PL: Repeat sequence families derived from mammalian tRNA genes. Nature 1985;317:819–822.
6 Deininger PL, Daniels GR: The recent evolution of mammalian repetitive DNA elements. Trends Genet 1986;2:76–80.
7 Ullu E, Tschudi C: Alu sequences are processed 7SL RNA genes. Nature 1984;312:171–172.
8 Xiong Y, Eickbush TH: Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 1990;9:3353–3362.
9 Inouye S, Yuki S, Saigo K: Sequence-specific insertion of the *Drosophila* transposable genetic element 17.6. Nature 1984;310:332–333.
10 Bannert N, Kurth R: Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci USA 2004;101(suppl 2):14572–14579.
11 Bureau TE, Wessler SR: Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 1992;4:1283–1294.
12 Feschotte C, Zhang X, Wessler SR: Miniature inverted-repeat transposable elements and their relationship to established DNA transposons; in Craig NL, Craigie R, Gellert M, Lambowitz A (eds): Mobile DNA II. (ASM Press, Washington, DC 2002).
13 Jordan IK: Evolutionary tinkering with transposable elements. Proc Natl Acad Sci USA 2006;103:7941–7942.
14 Cappello J, Handelsman K, Lodish HF: Sequence of *Dictyostelium* DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. Cell 1985;43: 105–115.
15 Arkhipova IR, Pyatkov KI, Meselson M, Evgen'ev MB: Retroelements containing introns in diverse invertebrate taxa. Nat Genet 2003;33:123–124.
16 Doolittle WF, Sapienza C: Selfish genes, the phenotype paradigm and genome evolution. Nature 1980;284:601–603.
17 Orgel LE, Crick FH: Selfish DNA: the ultimate parasite. Nature 1980;284:604–607.
18 Hickey DA: Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 1982;101:519–531.
19 Kidwell MG, Lisch DR: Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution Int J Org Evolution 2001;55:1–24.
20 Miller WJ, Hagemann S, Reiter E, Pinsker W: P-element homologous sequences are tandemly repeated in the genome of *Drosophila guanche*. Proc Natl Acad Sci USA 1992;89: 4018–4022.
21 Gould SJ, Vrba E: Exaptation – a missing term in the science of form. Paleobiology 1982;8:4–14.
22 Trelogan SA, Martin SL: Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. Proc Natl Acad Sci USA 1995;92: 1520–1524.
23 Robertson HM, Engels WR: Modified P elements that mimic the P cytotype in *Drosophila melanogaster*. Genetics 1989;123:815–824.
24 Britten RJ: DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci USA 1996;93:9374–9377.
25 Jordan IK, Rogozin IB, Glazko GV, Koonin EV: Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 2003;19:68–72.
26 van de Lagemaat LN, Landry JR, Mager DL, Medstrand P: Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 2003;19:530–536.
27 Blackburn EH: Structure and function of telomeres. Nature 1991;350:569–573.
28 Eickbush TH: Telomerase and retrotransposons: which came first? Science 1997;277:911–912.
29 Oettinger MA, Schatz DG, Gorka C, Baltimore D: RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science 1990;248:1517–1523.
30 Tonegawa S: Somatic generation of antibody diversity. Nature 1983;302:575–581.

31 Spanopoulou E, Zaitseva F, Wang FH, Santagata S, Baltimore D, Panayotou G: The homeodomain region of Rag-1 reveals the parallel mechanisms of bacterial and V(D)J recombination. Cell 1996;87:263–276.
32 Agrawal A, Eastman QM, Schatz DG: Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 1998;394:744–751.
33 Kapitonov VV, Jurka J: RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 2005;3:e181.
34 Cordaux R, Udit S, Batzer MA, Feschotte C: Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci USA 2006;103:8101–8106.
35 Robertson HM, Zumpano KL: Molecular evolution of an ancient mariner transposon, Hsmar1, in the human genome. Gene 1997;205:203–217.
36 Bundock P, Hooykaas P: An *Arabidopsis* hAT-like transposase is essential for plant development. Nature 2005;436:282–284.
37 Jurka J, Kapitonov VV: Sectorial mutagenesis by transposable elements. Genetica 1999;107: 239–248.
38 Kipling D, Warburton PE: Centromeres, CENP-B and Tigger too. Trends Genet 1997;13:141–145.
39 Smit AF: Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 1999;9:657–663.
40 Volff JN: Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 2006;28:913–922.
41 Britten R: Transposable elements have contributed to thousands of human proteins. Proc Natl Acad Sci USA 2006;103:1798–1803.
42 Nekrutenko A, Li WH: Transposable elements are found in a large number of human protein-coding genes. Trends Genet 2001;17:619–621.
43 Pavlicek A, Clay O, Bernardi G: Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data? FEBS Lett 2002;523:252–253.
44 Sorek R, Ast G, Graur D: Alu-containing exons are alternatively spliced. Genome Res 2002;12: 1060–1067.
45 Krull M, Brosius J, Schmitz J: Alu-SINE exonization: en route to protein-coding function. Mol Biol Evol 2005;22:1702–1711.
46 Nekrutenko A, Chung WY, Li WH: An evolutionary approach reveals a high protein-coding capacity of the human genome. Trends Genet 2003;19:306–310.
47 Wilson C, Goetting-Minesky P, Nekrutenko A: mNSC1 shows no evidence of protein-coding capacity. Gene 2006;370:83–85.
48 Claverie JM, Makalowski W: Alu alert. Nature 1994;371:752.
49 Gotea V, Makalowski W: Do transposable elements really contribute to proteomes? Trends Genet 2006;22:260–267.
50 Smit AFA, Hubley R, Green P: RepeatMasker Open-3.0. 1996–2004.
51 Chi N, Epstein JA: Getting your Pax straight: Pax proteins in development and disease. Trends Genet 2002;18:41–47.
52 Robson EJ, He SJ, Eccles MR: A PANorama of *PAX* genes in cancer and development. Nat Rev Cancer 2006;6:52–62.
53 Bopp D, Burri M, Baumgartner S, Frigerio G, Noll M: Conservation of a large protein domain in the segmentation gene paired and in functionally related genes of *Drosophila*. Cell 1986;47: 1033–1040.
54 Xu W, Rould MA, Jun S, Desplan C, Pabo CO: Crystal structure of a paired domain-DNA complex at 2.5 A resolution reveals structural basis for Pax developmental mutations. Cell 1995;80:639–650.
55 Breitling R, Gerber JK: Origin of the paired domain. Dev Genes Evol 2000;210:644–650.
56 Franz G, Loukeris TG, Dialektaki G, Thompson CR, Savakis C: Mobile Minos elements from *Drosophila hydei* encode a two-exon transposase with similarity to the paired DNA-binding domain. Proc Natl Acad Sci USA 1994;91:4746–4750.
57 Ivics Z, Izsvak Z, Minter A, Hackett PB: Identification of functional domains and evolution of Tc1-like transposable elements. Proc Natl Acad Sci USA 1996;93:5008–5013.

58 Hadrys T, DeSalle R, Sagasser S, Fischer N, Schierwater B: The Trichoplax *PaxB* gene: a putative Proto-*PaxA/B/C* gene predating the origin of nerve and sensory cells. Mol Biol Evol 2005;22: 1569–1578.
59 Vorobyov E, Horst J: Getting the Proto-Pax by the Tail. J Mol Evol 2006;63:153–164.
60 Zdobnov EM, Campillos M, Harrington ED, Torrents D, Bork P: Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. Nucleic Acids Res 2005;33: 946–954.

I. King Jordan
School of Biology
Georgia Institute of Technology
310 Ferst Drive
Atlanta, GA 30332-0230 (USA)
Tel. 404-385-2224, Fax 404-894-0519, E-Mail king.jordan@biology.gatech.edu

Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 163–174

Modulation of Host Genes by Mammalian Transposable Elements

W. Makałowski[a], *Y. Toda*[b]

[a]Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, University Park, Ill. USA; [b]Third Wave Japan, Tokyo, Japan

Abstract

Interspersed repetitive sequences are major components of eukaryotic genomes. They comprise about 50% of the mammalian genome. They interact with the whole genome and influence its evolution. They do this in many ways, e.g. by serving as recombination hotspots, providing a mechanism for genomic shuffling and a source of 'ready-to-use' motifs for new transcriptional regulatory elements, polyadenylation signals, and protein-coding sequences. In this review we discuss the consequences of exaptation of sequences originated in tansposable elements with focus on events that influence protein coding genes.

Copyright © 2007 S. Karger AG, Basel

The vast majority of a mammalian genome does not code for proteins and a large fraction of its DNA originated in transposable elements [1]. The late Susumu Ohno coined the term 'junk DNA' to describe this part of a genome [2]. Although catchy, the term 'junk DNA' for many years repelled mainstream researchers from studying noncoding DNA. Fortunately, more and more biologists regard repetitive elements as a genomic treasure [3, 4]. It appears that transposable elements are not useless DNA. They interact with the surrounding genomic environment and increase the ability of the organism to evolve. They do this in many ways, e.g. by serving as recombination hotspots, providing a mechanism for genomic shuffling and a source of 'ready-to-use' motifs for new transcriptional regulatory elements, polyadenylation signals, and protein-coding sequences. This review is divided into two parts. The first part describes briefly the structure of different classes of transposable elements that can be found in mammalian genomes. The second part discusses the consequences of

Table 1. Representation of transposable elements in different taxonomic groups

Class of elements	All Mammals	Primates	Rodentia	Cetartiodactyla	Carnivora
DNA transposons	178	139	60	56	57
Endogenous retroviruses	866	369	251	58	93
LTR retrotransposons	104	24	73	3	3
Non-LTR retrotransposons	357	158	107	57	51
SINEs	173	48	60	25	20
Total Interspersed Repeats	1,558	715	515	197	227

Data taken from issue 11.07 (August 2, 2006) of RepBase [47].

exaptation of sequences originated in transposable elements with focus on elements that influence protein coding genes.

Transposable Elements

Eukaryotic genomes contain interspersed repeat sequences that have largely amplified in copy number by movement throughout the genome. Transposable elements can be divided into two classes based on the mode of transposition: class I transposons or retrotransposons, and class II or DNA transposons [5]. All vertebrate genomes contain multiple families of transposable elements. Table 1 lists a number of transposable element families in different taxonomic groups of mammals.

Class I Transposons

Retrotransposons are mobile elements that multiply using a full length copy of their transcript intermediate that are reverse transcribed and reintegrated into a genome. Retrotransposons can be further divided into two major categories: LTR retrotransposons and non-LTR retrotransposons.

LTR retrotransposons are characterized by the presence of long terminal repeats (LTR) at both ends, and similarity to retroviruses. Both exogenous retroviruses and LTR retrotransposons contain a *gag* gene, that encodes a viral particle coat, and a *pol* gene that encodes a reverse transcriptase, ribonuclease H, and integrase, which provide the enzymatic machinery for reverse transcription and integration into the host genome. Reverse transcription occurs within the viral or viral-like particle (GAG) in the cytoplasm, and it is a multi-step process [6]. Unlike LTR retrotransposons, exogenous retroviruses contain an *env* gene, which encodes an envelope that facilitates their migration to other

cells. Some LTR retrotransposons may contain remnants of an *env* gene [7] but their insertion capabilities are limited to the originating genome. This would rather suggest that they originated in exogenous retroviruses by losing the *env* gene. However, there is evidence that suggests the contrary, given that LTR retrotransposons can acquire the *env* gene and become infectious entities [8]. Currently, most of the LTR sequences (85%) are found only as isolated LTRs, with the internal sequence being lost most likely due to homologous recombination between flanking LTRs [9]. It is interesting to note that LTR retrotransposons target their reinsertion to specific genomic sites, often around genes, with putative important functional implications for a host gene [7]. Lander et al. [9] estimate that 450,000 LTR copies make up about 8% of our genome.

Non-LTR retrotransposons miss the long terminal repeats, have a poly-A tail at the 3′ end, and are flanked by target site duplication (TSD). Several clades or superfamilies of non-LTR retrotransposons have been defined based on phylogenetic studies using the sequences of proteins encoded by each element.

The best represented non-LTR retrotransposons are **LINEs** (**L**ong **IN**terspersed **E**lements), which comprise about 21% of our genome (about 850,000 copies) [9]. Among these, the best described element is the LINE1 (L1) non-LTR retrotransposon. A full copy of L1 is about 6 kb long, contains a PolII promoter, and two ORFs. The function of the ORF1 protein is unclear, but it contains non-specific Zn-finger, leucine zipper, and coiled-coil motifs [10]. The second ORF encodes an endonuclease, which makes a single stranded nick in the genomic DNA, and a reverse transcriptase, which uses the nicked DNA to prime reverse transcription of LINE RNA from the 3′ end [11]. Reverse transcription is often unfinished, leaving behind fragmented copies of LINE elements, hence most of the L1-derived repeats are short, with an average size of 900 bp. L1 is the only LINE retroposon still active in the human genome [9]. Because they encode their own retrotransposition machinery, LINE elements are regarded as autonomous retrotransposons. In the human genome there are two other LINE-like repeats, L2 and L3, distantly related to L1. They are part of the CR1 clade, which has members in various metazoan species, including fruit fly, mosquito, zebrafish, pufferfish, turtle, and chicken [12].

SINEs, or **S**hort **IN**terspersed **E**lements, form another important group of non-LTR retrotransposons, evolved from RNA genes, such as 7SL, and tRNA genes [13]. By definition, they are short, up to 1,000 bp long. They do not encode their own retrotranscription machinery and are considered as non-autonomous elements. They are thought to be retrotransposed by the L1 machinery [14]. The outstanding member of this class from the human genome is the Alu repeat, which contains a cleavage site for the *Alu*I restriction enzyme, where it got its name from [15]. Alus are primate specific elements, but they have a rodent relative, B1 element, which is a monomeric repeat, also derived

from the 7SL RNA gene. MIRs, by contrast, spread before eutherian radiation, and their copies can be found in different mammalian groups including marsupials and monotremes [16].

SVA elements are composite repetitive entities named after their main components: SINE, VNTR and Alu [17]. Usually, they contain the hallmarks of the retroposition, i.e. they are flanked by TSDs and terminated by a poly(A) tail. It seems that SVA elements are non-autonomous retrotransposons mobilized by L1 machinery and they are thought to be transcribed by RNA polymerase II. SVAs are transpositionally active causing some human diseases [18]. They originated less than 25 million years ago and they form the youngest retrotransposon family with about 3,000 copies in the human genome [19]. Interestingly, similarly to L1 elements, they can transduce downstream sequences during their movement. It has been shown recently that about 53 kb of the human genomic sequences has been duplicated by SVA-mediated transductions, including three independent duplications of the entire *AMAC* gene [20].

Retro(pseudo)genes are products of reverse transcription of a spliced (mature) mRNA. Hence, their characteristic features are: an absence of both 5′-promoter sequence and introns, the presence of flanking direct repeats and a 3′end-polyadenosine tract [21]. Retro(pseudo)genes like other retrotransposons have been inserted into the genome as double-stranded sequence generated from a single-stranded RNA. Processed pseudogenes, as sometimes retropseudogenes are called, have been generated in vitro at a low frequency in human HeLa cells via mRNA from a reporter gene [22]. The source of the reverse transcription machinery in humans and other vertebrates seems to be active L1 elements [23]. However, not all retroposed messages have to end up as pseudogenes. About 20% of mammalian protein encoding genes lack introns in their ORFs [24]. It is conceivable that many if not all genes lacking introns arose by retroposition. Some genes are known to be very prolific in producing retroelements. For instance, in the human genome there are over 2,000 retropseudogenes for ribosomal proteins [25]. A recent genome-wide study showed that the human genome harbors about 20,000 pseudogenes, 72% of which arose through retroposition [26]. Interestingly, vast majority (92%) of them are quite recent transpositions that occurred after primate/rodent divergence [24].

Class II elements

Class II elements move by a conservative cut-and-paste mechanism, the excision of the donor element is followed by its reinsertion elsewhere in the genome. **DNA transposons** are abundant in bacteria, where they are called insertion sequences (IS), but are present in many other genomes, including Metazoa [7]. They are characterized by terminal inverted repeats, and encode a transposase that binds near the inverted repeats and mediates mobility through

a ‘cut and paste’ mechanism [27]. This process is not usually a replicative one, unless the gap caused by excision is repaired using the sister chromatid. When inserted at a new location, the transposon is flanked by small gaps, which, when filled by host enzymes, cause duplication of the sequence at the target site. These are called target site duplications (TSDs), and their length is characteristic for particular transposons. Based on sequence similarity of the transposase, eukaryotic DNA transposons fall into two classes: the *Ac/hobo* class, characterized by 8 bp TSD, and the *Tc1/mariner* class, characterized by TA dimer duplication [28]. Members of both classes have been found in vertebrate genomes [28, 9], however it seems that there are no active DNA transposons in mammalian genomes. It is estimated that DNA transposons make up about 3% of the human genome. The most abundant members of this class are *mariner* elements that comprise about 1.5% of the human genome.

Exaptation of Sequences Originated in Transposable Elements (TEs)

Exaptation is a term introduced by Gould and Vrba [29] to explain how different characters may be adopted to new roles regardless of their original function or no function at all. Those characters might have been shaped by natural selection for specific functions, or might have had no function at all. The concept was applied to genomic level by Brosius and Gould [30] and fits perfectly here, as TE-derived sequences were originally part of mobile TEs and played a different (if any) role.

TE Cassettes in mRNAs

TE cassettes from different types of elements can be found surprisingly frequently in mammalian messages (table 2). Fractions of mammalian mRNAs with TE cassettes vary from $<1\%$ in bovine to about 15% in humans. However, these differences may reflect both our knowledge of transposable elements in a given species and annotation level of protein coding genes, especially their alternative forms. TE cassettes can be found in all three regions of an mRNA, i.e. 5′ UTR, CDS, and 3′ UTR. Several special cases are particularly interesting: TE cassettes at either end of a message and these at the border of coding sequences and UTRs. A TE cassette at the very 5′ end suggests that promoter signals are donated by a TE sequence. Such a cassette at 3′ end suggests that a polyadenylation signal actually lies within a TE. A TE cassette at the beginning or the end of coding sequence strongly suggests that a TE provided a START or STOP codon, respectively.

The exonized transposable element was first noticed in a disease phenotype. A single point mutation in an Alu element residing in the third intron of

Table 2. Mammalian mRNAs with TE cassettes

Species	Number of analyzed transcripts[a]	5′ UTR[b]	CDS[c]	3′ UTR[d]
Bos taurus	32,807 (2,730; 4,504)	19 (3)	170 (34; 46)	33 (3)
Canis familiaris	30,229 (6,354; 6,716)	8 (2)	503 (24; 188)	39 (13)
Homo sapiens	40,078 (24,003; 24,691)	1,428 (383)	1,854 (649; 901)	3,706 (1,224)
Macaca mulatta	34,901 (14,907; 13,920)	260 (50)	1,619 (482; 835)	1,434 (489)
Mus musculus	30,007 (17,787; 18,222)	436 (64)	553 (233; 329)	2,720 (410)
Pan troglodytes	35,864 (17,304; 16,637)	694 (164)	828 (287; 424)	4,056 (1,188)
Rattus norvegicus	33,071 (10,778; 10,642)	83 (10)	1,135 (611; 589)	1,088 (279)

[a]Total number of transcripts with annotated CDS; numbers in parentheses refer to messages with annotated 5′ UTR and 3′ UTR, respectively.

[b]Total number of 5′ UTRs with a TE cassette; numbers in parentheses refer to UTRs with a TE cassette that overlap with the begining of the transcript.

[c]Total number of CDSs with a TE cassette; numbers in parentheses refer to CDSs with a TE cassette overlapping with the START and STOP codons, respectively.

[d]Total number of 3′ UTRs with a TE cassette; numbers in parentheses refer to 3′ UTRs with a TE cassette that overlap with the end of the transcript and likely providing polyadenylation signal to such a transcript.

Mammalian transcriptoms were downloaded from the Ensembl database using BioMart server [48] and searched for the TEs using RepeatMasker.

the human ornithine aminotransferase gene activated cryptic splicing sites and consequently lead to introduction of a partial Alu element into an open reading frame [31]. An in-frame STOP codon carried by an Alu cassette caused a truncated protein and ornithine δ-aminotransferase deficiency was observed, too. This discovery led to the hypothesis that similar mechanisms are used for fast evolutionary changes in protein structure leading to increased protein variability [32]. Recently, Krull et al. investigated the details of the Alu exonization process [33]. It appears that exonization is a multistep process that involves: (a) integration of an Alu within a protein coding gene locus (usually within one of the introns), (b) mutational modulation of the sequence that creates or activates cryptic splicing sites and maintenance of an open reading frame. These steps are often separated in time by millions of years and can be reversed [33].

The big mystery is how a genome adapts to the drastic changes induced to a protein by the insertion of a mobile element into the coding region of its gene. Two studies demonstrated how this process takes place without disturbing the function of the original protein [34, 35]. One way is to keep a TE cassette as an alternatively spliced exon (fig. 1). However, this requires maintaining the delicate balance of signals that cause an exon to be spliced alternatively – too strong signal causes an exon to be spliced constitutively but too weak signal

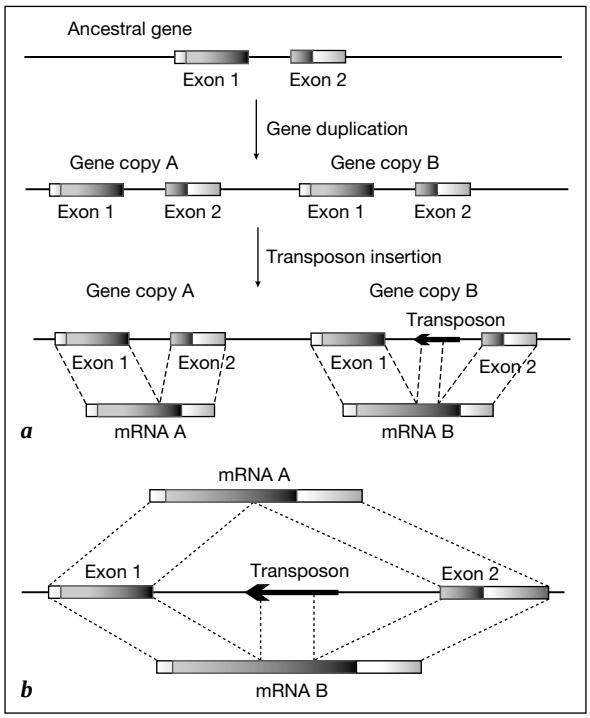


Fig. 1. Two ways of exapting a TE cassette to the coding region of a gene without destroying the gene's function. ***a*** Insertion of a TE cassette is preceded by a gene duplication. ***b*** A TE cassette is inserted into the mRNA as an alternative exon. In both cases, the genome gains two forms of the transcript – one with and one without a TE cassette.

leads to exon skipping. Another way to exploit an evolutionary novelty without disturbing the function of the original protein is gene duplication. Gene duplication is one of the major ways in which organisms can generate new genes [36]. After a gene duplication, one copy maintains its original function whereas the other is free to evolve and can be used for 'nature's experiments'. Usually, this is accomplished through point mutations and the whole process is very slow. However, recycling some modules that already exist in a genome (for example, in transposons) can speed up the natural mutagenesis process tremendously. This is exactly how bovine BCNT protein gained its endonuclease domain – capturing it from the ruminant retrotransposable element-1 (RTE-1). Bovines, unlike humans and mice, have two copies of the *BCNT* gene (also called *CFDP1*), one 'classical' similar to other mammalian *CFDP1* genes, and another one with the endonuclease domain [34].

TE cassettes in mammalian ORFs are not uncommon [37]. This is in contrast to a small number of TE cassettes found in functional proteins (0.1% vs. 4%) [38]. A few known examples of TE cassettes found in functional proteins originated in older repeats, i.e. L3 or MIR [38]. Interestingly, most TE cassettes at the transcript level are derived from young TEs, and appear in a minor, alternatively spliced form of cognate mRNAs [37]. They may even persist as such over long evolutionary periods [33], which may indicate that they do not represent successful exaptations for protein coding purposes, nor do they represent intermediate stages of such events. They rather play a different important role, or otherwise they would be lost. Interestingly, Oh et al. showed that coexpression of wild type human epithelial sodium channel (hENaC) α, β, and γ subunits with an Alu-containing splicing variant of the α subunit (hαENaC+Alu) enhanced the expression of the amiloride-sensitive in oocytes [39]. The significant number of TE-containing transcripts may indicate that the role of TEs in regulation of gene expression and function is more important than currently acknowledged.

TE Sequences in Promoter Regions

One of the most direct influences of transposable elements on the host genome is their role in modulating the structure and expression of 'resident' genes. After the discovery that long terminal repeats carry promoter and enhancer motifs it became clear that integration of such elements in proximity of a host gene must have an influence on this gene expression [40]. It seems that a sizable fraction of eukaryotic, gene-associated regulatory elements arose in this modular fashion by insertion of TEs, and not only by point mutations of static neighboring sequences. When a TE is inserted upstream from a gene, a few short motifs can be conserved if they were subjected to selective pressure as promoters or enhancers of transcription. Even though the rest of the TE sequence might evolve beyond recognition due to absence of functional constraints, TEs are hence exapted into a novel function [30]. A recent survey that analyzed 846 functionally characterized cis-regulatory elements from 288 genes showed that 21 of those elements (~2.5%) from 13 genes (~4.5%) reside in TE-derived sequences [41]. The same study showed that TE-derived sequences are present in many more (~24%) promoter regions, defined as ~500 bp located upstream of functionally characterized transcription initiation site. Similarly, van de Lagemaat et al. showed that the 5′ UTRs of a large proportion of mammalian mRNAs contain TE fragments, suggesting that they play a role in regulation of gene expression [42].

Another recent study demonstrated high potential of transposable elements in regulation of gene expression [43]. It appears that all 20 investigated binding site classes are over-represented in at least one TE class with three (helix-loop-helix, TATA binding proteins, runt) being over-represented in all four TE classes. These are all transcription factor binding sites that control

the expression of many genes, while binding sites over-represented in only one of the TE classes, RF-X, zinc finger/GATA, and heteromeric CCAAT factors/histone folds, appear to have more specific functions. Interestingly, there is a reversed correlation between TE sequence abundance in promoter regions and their capacity of carrying regulatory signals, i.e. the most abundant in promoter TEs (SINEs) carry the least number of protein binding sites. This study clearly demonstrates that occupying half of the human genome, TEs have a big potential to influence gene regulation at genomic scale by carrying potential transcription regulating signals [43].

Transposable Elements and microRNAs

MicroRNAs are small (about 20 nt) RNAs that educe mRNA degradation if they bind perfectly to the target mRNA, or arrest mRNA translation if binding is not perfect. They are processed from larger, up to 100 nt long, precursors. Recent work of Smalheiser and Torvick shed some light on the TE involvement in the whole process [44]. First, they demonstrated that eight mammalian microRNAs originated in TEs. Interestingly, four of them derived from a LINE-2 element. Two of these perfectly match a large family of transcripts that contain L2 elements in their 3′ UTRs [44]. This study suggests that the insertion of a TE into a new genomic location may create a new microRNA gene during mammalian evolution. They may also provide lineage specific microRNAs, for instance L2-derived mir-95 seems to be human (or primate) specific. Interestingly, although mir-151 is conserved in primates and rodents, its target mRNAs are human specific because the L2 insertion into 3′ UTR has occurred de novo during primate evolution.

The same authors also analyzed Alu elements residing in 3′ UTRs to see if they can serve as targets for microRNAs [45]. They found that almost 30 human microRNAs have potential targets within specific Alu elements inserted in 3′ UTRs. Scanning Alu elements inserted into 3′ UTRs in sense orientation revealed that 83% of these sequences contain potential targets for microRNAs [45]. This study shows a great potential of Alu elements to initiate mRNA degradation via microRNA mechanism. Since Alu elements are not the only TEs present in mammalian 3′ UTRs it will be extremely interesting to see if other transposons exhibit similar potential.

Final Remarks

For a long time transposable elements have been regarded as genomic parasites or at least selfish, useless elements. However, almost every day we discover new instances demonstrating TE involvement in basic cellular processes,

especially their role in shaping mammalian genomes is invaluable. TEs are responsible for segmental duplications, origination of new genetic material, shuffling the old one, and many other events. Risking personification of biological processes, we can say that evolution is too wise to waste this valuable information. Therefore, TE-derived DNA should not be called junk DNA but a genomic scrap yard, because it serves a reservoir of ready-to-use segments for nature's evolutionary experiments [46, 3].

References

1 Makalowski W: The human genome structure and organization. Acta Biochim Pol 2001;48: 587–598.
2 Ohno S: So much "junk" DNA in our genome; in Smith HH (ed): Brookhaven Symposia in Biology. Gordon & Breach, New York, 1972, No. 23, pp 366–370 .
3 Makalowski W: Genomic scrap yard: how genomes utilize all that junk. Gene 2000;259:61–67.
4 Volff JN: Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 2006;28:913–922.
5 Finnegan DJ: Eukaryotic transposable elements and genome evolution. Trends Genet 1989;5: 103–107.
6 Voytas D, Boeke JD: Ty1 and Ty5 of *Saccharomyces cerevisiae*, in Craig N, et al. (eds): Mobile DNA II. ASM Press, Washington, DC, 2002.
7 Kazazian HH Jr.: Mobile elements: drivers of genome evolution. Science 2004;303:1626–1632.
8 Malik HS, Henikoff S, Eickbush TH: Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 2000;10:1307–1318.
9 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al: Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. Nature 2001;409:860–921.
10 Poulter R, Butler M, Ormandy J: A LINE element from the pufferfish (fugu) *Fugu rubripes* which shows similarity to the CR1 family of non-LTR retrotransposons. Gene 1999;227:169–179.
11 Ostertag EM, Kazazian HH Jr.: Biology of mammalian L1 retrotransposons. Annu Rev Genet 2001;35:501–538.
12 Kapitonov VV, Jurka J: A novel class of SINE elements derived from 5S rRNA. Mol Biol Evol 2003;20:694–702.
13 Ohshima K, Okada N: SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. Cytogenet Genome Res 2005;110:475–490.
14 Dewannieux M, Esnault C, Heidmann T: LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 2003;35:41–48.
15 Houck CM, Rinehart FP, Schmid CW: A ubiquitous family of repeated DNA sequences in the human genome. J Mol Biol 1979;132:289–306.
16 Jurka J, Zietkiewicz E, Labuda D: Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. Nucleic Acids Res 1995;23:170–175.
17 Shen L, Wu LC, Sanlioglu S, Chen R, Mendoza AR, et al: Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J Biol Chem 1994;269:8466–8476.
18 Ostertag EM, Goodier JL, Zhang Y, Kazazian HH Jr.: SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 2003;73:1444–1451.
19 Wang H, Xing J, Grover D, Hedges DJ, Han K, et al: SVA elements: a hominid-specific retroposon family. J Mol Biol 2005;354:994–1007.
20 Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA: From the cover: eukaryotic transposable elements and genome evolution special feature: emergence of primate genes by retrotransposon-mediated sequence transduction. Proc Natl Acad Sci USA 2006;103:17608–17613.

21 Vanin EF: Processed pseudogenes: characteristics and evolution. Annu Rev Genet 1985;19: 253–272.
22 Maestre J, Tchenio T, Dhellin O, Heidmann T: mRNA retroposition in human cells: processed pseudogene formation. EMBO J 1995;14:6333–6338.
23 Esnault C, Maestre J, Heidmann T: Human LINE retrotransposons generate processed pseudogenes. Nat Genet 2000;24:363–367.
24 Sakharkar MK, Kangueane P, Petrov DA, Kolaskar AS, Subbiah S: SEGE: a database on 'intron less/single exonic' genes from eukaryotes. Bioinformatics 2002;18:1266–1267.
25 Zhang Z, Harrison P, Gerstein M: Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res 2002;12:1466–1482.
26 Torrents D, Suyama M, Zdobnov E, Bork P: A genome-wide survey of human pseudogenes. Genome Res 2003;13:2559–2567.
27 Mizuuchi K, Baker T: Chemical mechanisms for mobilizing DNA; in Craig N, et al. (eds): Mobile DNA II. ASM Press, Washington, DC, 2002.
28 Smit AF: The origin of interspersed repeats in the human genome. Curr Opin Genet Dev 1996;6: 743–748.
29 Gould SJ, Vrba ES: Exaptation – a missing term in the science of form. Paleobiology 1982;8: 4–15.
30 Brosius J, Gould SJ: On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proc Natl Acad Sci USA 1992;89:10706–10710.
31 Mitchell GA, Labuda D, Fontaine G, Saudubray JM, Bonnefont JP, et al: Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. Proc Natl Acad Sci USA 1991;88:815–819.
32 Makalowski W, Mitchell GA, Labuda D: Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet 1994;10:188–193.
33 Krull M, Brosius J, Schmitz J: Alu-SINE exonization: en route to protein-coding function. Mol Biol Evol 2005;22:1702–1711.
34 Iwashita S, Osada N, Itoh T, Sezaki M, Oshima K, et al: A transposable element-mediated gene divergence that directly produces a novel type bovine bcnt protein including the endonuclease domain of RTE-1. Mol Biol Evol 2003;20:1556–1563.
35 Lev-Maor G, Sorek R, Shomron N, Ast G: The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. Science 2003;300:1288–1291.
36 Ohno S: Evolution by Gene Duplication. Springer-Verlag, New York 1970.
37 Lorenc A, Makalowski W: Transposable elements and vertebrate protein diversity. Genetica 2003;118:183–191.
38 Gotea V, Makalowski W: Do transposable elements really contribute to proteomes? Trends Genet 2006;22:260–267.
39 Oh YS, Lee S, Won C, Warnock DG: An Alu cassette in the human epithelial sodium channel. Biochim Biophys Acta 2001;1520:94–98.
40 Sverdlov ED: Perpetually mobile footprints of ancient infections in human genome. FEBS Lett 1998;428:1–6.
41 Jordan IK, Rogozin IB, Glazko GV, Koonin EV: Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 2003;19:68–72.
42 van de Lagemaat LN, Landry JR, Mager DL, Medstrand P: Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 2003;19:530–536.
43 Thornburg BG, Gotea V, Makalowski W: Transposable elements as a significant source of transcription regulating signals. Gene 2006;365:104–110.
44 Smalheiser NR, Torvik VI: Mammalian microRNAs derived from genomic repeats. Trends Genet 2005;21:322–326.
45 Smalheiser NR, Torvik VI: Alu elements within human mRNAs are probable microRNA targets. Trends Genet 2006;22:532–536.
46 Makalowski W: SINEs as a genomic scrap yard: an essay on genomic evolution; in Maraia RJ (ed): The Impact of Short Interspersed Elements (SINEs) on the Hpst Genome. R.G. Landes Company, Austin, 1995, pp 81–104.

47 Jurka J: Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 2000;16:418–420.
48 Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, et al: BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 2005;21: 3439–3440.

Wojciech Makałowski
Institute of Bioinformatics
University of Muenster, Von-Esmarch-Straße 54
D-48149 Muenster (Germany)
E-Mail: wojmak@uni-muenster.de

Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 175–190

Modern Genomes with Retro-Look: Retrotransposed Elements, Retroposition and the Origin of New Genes

J.-N. Volff[a], *J. Brosius*[b]

[a]Institut de Génomique Fonctionnelle, INRA, CNRS, Université Lyon 1, Ecole Normale Supérieure, Lyon, France; [b]Institute of Experimental Pathology, Molecular Neurobiology (ZMBE), University of Münster, Münster, Germany

Abstract

A fascinating evolutionary facet of retroposition is its ability to generate a dynamic reservoir of sequences for the formation of new genes within genomes. Retroelement genes, such as *gag* from retrotransposons or envelope genes from endogenous retroviruses, have been repeatedly exapted and domesticated during evolution. Such genes fulfill now useful novel functions in diverse aspects of host biology, for example placenta formation in mammals. New protein-coding genes can also be generated through the reverse transcription of mRNA from 'classical' genes by the enzymatic machinery of autonomous retroelements. Many of these retrogenes, which generally show a modified expression pattern compared to their molecular progenitor, have a testis-biased expression and a potential role in spermatogenesis in different animals. New non-protein-coding RNA genes have also been repeatedly generated through retroposition during evolution. A striking evolutionary parallel has been observed between two such RNA genes, the rodent BC1 and the primate BC200 genes. Although both genes are derived from different types of sequences (tRNA and Alu short interspersed element, respectively), they are both expressed almost specifically in neurons, transported into the dendrites and included in ribonucleoprotein complexes containing the poly(A)-binding protein PABP. Both BC1 and BC200 RNA are able to inhibit translation in vitro and are progenitors of new families of short interspersed elements. These genes, which might play a role in animal behavior, provide an astonishing example of evolutionary convergence in two distinct mammalian lineages, which is also observed for placenta genes derived from endogenous retroviruses. Finally, there are indications that genes for small nucleolar RNAs (snoRNAs) and possibly microRNAs (miRNAs) can also be duplicated via retroposition. Taken together, these observations definitely demonstrate the major role of retroposition as mediator of genomic plasticity and contributor to gene novelties. Therefore, the 'retro-look' of genomes is in fact indicative of their modernity.

Copyright © 2007 S. Karger AG, Basel

Two major categories of discernible sequences cohabit within genomes: 'classical' genes and transposed elements (TEs). While the useful role of classical genes has been taken for granted from early on, the status of TEs has been more controversial. Once denigrated as purely selfish and parasitic [1], transposed elements have been gradually rehabilitated by a wave of new knowledge demonstrating their important role in the structure, function and evolution of genes and genomes [2, 3].

Given the tens of thousands of genes and hundreds of thousands up to millions of TEs generally present within genomes, the absence of genetic exchanges between both types of sequences would be surprising. Accordingly, it has been observed that transposed elements and viruses can capture cellular genes and hijack their function for their own purpose. On the other hand, many host genes have recruited TE sequences during evolution and use them for either their regulation or their coding potential (exonisation) [4–8]. These observations indicate that TEs serve as an evolutionary reservoir of sequences for genetic innovation.

This review will focus on another particularly important facet of the evolutionary impact of TEs and transposition: the generation of new protein-coding and RNA genes. More emphasis will be given to retrotransposed elements (also called retroelements, RTEs) and retroposition, a mechanism of transposition requiring the reverse transcription of an RNA intermediate. The enzyme involved, called reverse transcriptase, is able to copy RNA into cDNA for the production of new copies of the retroelement [9]. Hence, retroposition corresponds to a mechanism of DNA duplication (retroduplication).

Autonomous RTEs encode at least a reverse transcriptase and are classified into groups according to their structural features. Some RTEs, including the Ty1/copia and Ty3/gypsy retrotransposons as well as all known retroviruses, are delimited by long terminal repeats (LTRs). Another large group of RTEs, the non-LTR retrotransposons or Long INterspersed Elements (LINEs), is exemplified by the L1 elements particularly well represented in mammalian genomes. Non-autonomous sequences not coding for retroposition proteins, including Short INterspersed Elements (SINEs) and even classical genes, can be retroposed in trans by the enzymatic machinery of autonomous retroelements.

DNA transposons, the second major category of TEs, transpose through a mechanism without reverse transcription generally involving a transposase. Although numerous examples of genes derived from DNA transposons are known [reviewed in ref. 10], these elements are not object of the present review.

Protein-Coding Genes Derived from Retroelements

There is evidence that RTE genes have been repeatedly exapted and domesticated during evolution to fulfill new functions useful to the host [10].

Exapted protein-coding genes are essentially recognizable by the facts that (i) they have lost sequences necessary for retrotransposition but (ii) have conserved a coding potential and frequently evolve under purifying (negative) selection and (iii) they are present at orthologous positions in different species and have a low copy number.

Different genes encoded by LTR retrotransposons or retroviruses have been domesticated during evolution. This is the case for *gag* genes, which normally encode a structural protein required for the assembly of RNA molecules into cytoplasmic RTE particles. Gag proteins generally have one or several zinc finger domains. In RTE with LTRs, *gag* generally partially overlaps with a larger downstream open reading frame called *pol*, which encodes a polyprotein with protease, reverse transcriptase and integrase domains. A Gag-Pol fusion protein is produced by translational frameshift between *gag* and *pol*.

As many as 15–20 *gag*-derived genes originating from two independent families, *Mart* (at least 10 genes) and *Ma* (six genes in human), are present in mammalian genomes. *Mart* genes have been apparently derived from the *gag* gene(s) of a Ty3/gypsy LTR retrotransposon of the *Sushi* family, a type of RTE active in fish but not in mammals [11–13]. *Ma* genes are derived either from LTR retrotransposons or from retroviruses [14]. Interestingly, there are striking similarities between the *Mart* and *Ma* gene families. For example, both *Mart* and *Ma* genes are preferentially localised on the X chromosome, this possibly reflecting a mechanism of gene family expansion through serial intrachromosomal duplications. In both families, genes with or without translational frameshift and encoding protein products with or without Gag-like zinc finger are observed. At least one member of each family is involved in apoptosis [15, 16]. Interestingly, *Mart2/Peg10* knockout mice show early embryonic lethality due to defects in placenta formation [17]. This demonstrates that at least one *gag*-derived gene has an essential role in mammalian development. Two autosomal *Mart* genes are imprinted and paternally expressed [18]. This epigenetic regulation might be derived from a control mechanism against the ancestral retrotransposon. Finally, another unrelated *gag*-derived gene called *Fv1* (Friend virus susceptibility 1) encodes a protein that restricts murine leukemia virus replication in the mouse [19], and many additional candidates for domesticated *gag* genes have been recently identified in the human genome [20].

The *pol* gene of LTR retrotransposons and retroviruses encodes among others an integrase required for integration of the double-stranded cDNA after reverse transcription. Examples of exapted integrase genes include *Gin-1* in mammals [21] and possibly the yeast gene *Fob1p*, which encodes a protein regulating the number and recombination rate of ribosomal RNA genes [22].

Compared to classical LTR retrotransposons, retroviruses additionally encode an envelope (Env) glycoprotein that recognises membrane receptors of

the host cell and initiate the process of infection via virus/cell fusion. Intact *Env* genes are frequently found in genomes. They are generally carried by so-called endogenous retroviruses (ERVs), which correspond to integrated defective copies of retroviruses that have been introduced by infection into the germ line of their host millions of years ago. In primates, two exapted retrovirus *Env* genes called *Syncytin-1* and *-2* are involved in the formation of placenta, the nutritional and protective interface between mother and developing foetus [23, 24]. Both Syncytin proteins are placenta-specific and can promote cell fusion in vitro. The results obtained are consistent with a role in the fusion of trophoblast cells that leads to the formation of the syncytiotrophoblast layer, a continuous structure with microvillar surfaces facilitating exchanges between mother and foetus [23, 24]. Two *Env*-like Syncytin genes, *Syncytin-A* and *-B*, with functions potentially similar to human Syncytins, have also been identified in mouse [25]. Both genes have been introduced through infection approximately 20 MYA into the murine lineage. In sheep, the envelope of endogenous Jaagsiekte retroviruses regulates trophectoderm growth and differentiation in the periimplantation ovine conceptus [26]. Since primate, mouse and sheep genes are not orthologous, these observations provide strong evidence of convergent domestication in placental mammals. Other *Env* genes with protein-coding potential are present in mammalian genomes, but their cellular functions remain to be determined. *Env* genes have also been exapted independently in *Drosophila* and mosquitoes [27].

Finally, the enzyme telomerase, a ribonucleoprotein containing an RNA subunit serving as template for telomere replication of eukaryotic chromosomes, might correspond to a reverse transcriptase exapted from a non-LTR retrotransposon or from another type of RTE [28, 29].

Protein-Coding Retrogenes

Retrogenes are formed through reverse transcription of spliced mRNA transcripts and integration of the cDNA into the genome. This process, also called retroduplication, can generate additional copies of 'classical' genes. Compared to the progenitor gene, retrogenes are generally intronless. However, often introns are being formed in the 5′ or 3′ untranslated regions [30]. Retrogenes also have a poly(A) tail at their 3′ end and are flanked by short target site duplications if the event is more recent. Since this mechanism of gene duplication involves an mRNA intermediate, the original promoter is generally not included in the duplicated region. Therefore, in order to be expressed, retrogenes require to be integrated at the proximity of regulatory sequences of other genes, or close to other types of sequences with promoter activity, for example

transposed elements. Alternatively, if a gene has different promoters, transcription from a distal promoter might lead to the formation of an mRNA molecule including a proximal promoter, which might be able to drive the expression of the retrogene after retroposition [31].

After retroposition, most retrogenes degenerate as processed pseudogenes. However, retroposed sequences can also give rise to new functional copies of a progenitor gene, generally with a modified expression pattern [31, 32]. About 3,600 retrocopies have been identified in the human genome, one third of them being transcribed. Among them, at least 120 sequences correspond to bona fide genes [31]. Interestingly, in animals as different as mammals and insects, many protein-coding retrogenes are autosomal but originate from X-linked progenitor genes and have developed testis-biased expression [31–34]. This phenomenon is believed to compensate for X-chromosome silencing during male meiosis. In the mouse, one such retrogene, integrated in an intron of an autosomal gene and using its promoter and 5′ untranslated exon, has been shown to be essential for spermatogenesis [35].

RNA Genes

Initially, the idea was absurd that retroposition, a mechanism responsible for littering genomes with retropseudogenes, repetitive elements, such as SINEs and LINEs, and endogenous retroviruses would generate something useful, especially non-protein-coding RNA molecules, a class of macromolecules that are fossils from the by-gone RNA and RNP worlds and on their way out. Fortunately, these preconceptions have changed in that retroposition is being recognised as important mediator of genomic plasticity and contributor to genic novelties [36, 37] and the functional significance of non-protein-coding RNA (npcRNA) is finally being appreciated [38–40].

Neuronal BC1 RNA in Rodents

BC1 RNA was the first brain-specific npcRNA discovered. Initially, it was thought that BC1 RNA was a transcriptional by-product of ID repetitive elements [41] that belonged to the family of tRNA-related SINEs. Subsequently, it was established that BC1 RNA was generated by retroposition of $tRNA^{Ala}$ that yielded, instead of an inactive tRNA pseudogene, a tissue-specific single-copy gene encoding an npcRNA located between *Fgf15* and *Oraov1* on mouse chromosome 7 [42]. BC1 RNA (~150 nt) is twice as long as $tRNA^{Ala}$; the second half of the molecule is contributed by a retroposition-related adenosine-rich

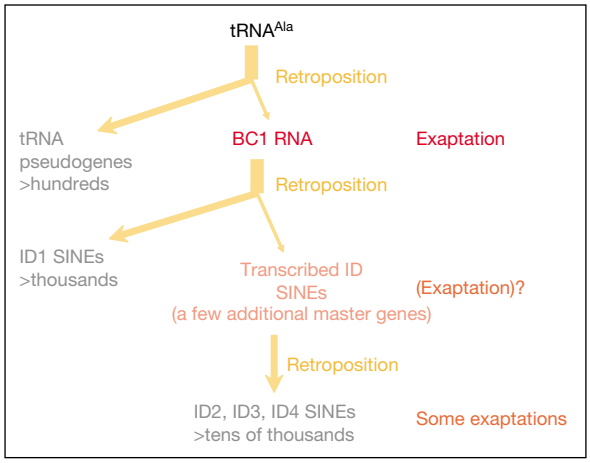


Fig. 1. Biogenesis of BC1 RNA and derived retrosequences. Retroposition of a small non-protein-coding RNA (npcRNA) often leads to a sizeable number of retrosequences including SINEs. Only a minority of these retroposed sequences are recruited or exapted into a function as it is the case for neural dendritic BC1 RNA. BC1 RNA, in turn, is a more efficient template for retroposition than its tRNA parent yielding thousands of SINEs of the ID1 type. Furthermore, one or a few of the ~10,000 ID1 elements must be transcribed in the rat as they became the master gene for tens of thousands of additional ID elements (ID2–4). It is likely that these are chance products of transcription without functional recruitment, at least as of yet. Of course, non-transcribed ID elements or those that are co-transcribed with an hnRNA or part of an mRNA (in their 3′ UTRs or as part of an exon) may have been exapted.

region in the centre as well as a non-repetitive region at the 3′ end acquired from the locus of integration. BC1 RNA is the parent gene of a subclass of ID SINEs [43], while certain transcribed ID elements became the masters of additional ID subfamilies [44].

The genesis of BC1 RNA and related SINEs is depicted in figure 1. BC1 RNA is transcribed by RNA polymerase III [45]. Apart from low level expression in pre-meiotic spermatogonia, the macromolecule is only found in neurons and there it is transported into dendrites including their distal processes [46]. The determinants (spatial codes) for dendritic transport reside in the tRNA-related 5′ domain that no longer folds into a cloverleaf structure but, instead, forms an extended stem loop [47]. A single bulged Uracil (position 22) is essential for dendritic transport; furthermore, a GA kink-turn motif in the apical part of the stem is important for distal dendritic delivery [48]. BC1 RNA also is developmentally regulated, the onset in neurons coincides with synapse formation [49] but is deregulated in immortalised cell cultures and certain tumours [50].

In evolutionary terms, BC1 RNA is relatively young and probably arose in the common ancestor of all rodents. By phylogenetic analysis, the BC1 RNA coding region is conserved at significantly higher levels than the flanking regions, pointing to a selective advantage for its conservation in rodents. The RNA is complexed with proteins as a ribonucleoprotein (RNP). About a dozen of proteins have been suggested as RNP components; only one is consistent, namely the poly(A)-binding protein PABP [51]. In in vitro translation assays using rabbit reticulocyte lysate as well as in transfected cell cultures, naked BC1 RNA inhibits translation of any reporter mRNA [52]. The adenosine-rich region was identified as the RNA domain responsible for the inhibitory effects. Consistently, binding of PABP prior to addition to the translation assay also has a much milder effect on translation, indicating that most of the outcome might be mediated by competition for PABP [52], an important translation initiation factor. Nevertheless, in the cell and especially in dendritic post-synaptic microdomains, modulation of distribution of translation factors such as PABP involving dendritic mRNAs and BC1 RNP, perhaps along with miRNAs, might play a role in regulation of post-synaptic protein biosynthesis, a mechanism thought to underlie synaptic plasticity including learning and memory [53]. Superficially, it is somewhat unexpected then, that deletion of the gene encoding BC1 RNA did not lead to any detectable deficiencies in learning and memory. Instead, a reduced exploratory behavior, possibly mediated by higher levels of anxiety, was observed in mice devoid of BC1 RNA [54]. The underlying biochemical pathways that are responsible for this behavioral change await identification.

Once more, it would have been surprising to find a gene product that is restricted to a single, albeit large and successful, mammalian order to be solely responsible for vital functions such as memory and learning, functions that are not only important to the survival of mammals but also of all vertebrates and many invertebrates.

Alu-Derived Neuronal npcRNAs in Primates

It is interesting to note that primates express an analogous but evolutionarily unrelated RNA, that might function in a similar manner as BC1 RNA in rodents. As it happened, neuron-specific, dendritic BC200 RNA arose from a monomeric Alu SINE element in a common ancestor of Anthropoidea [55, 56]. Initially monomeric, Alu elements (B1 in Glires), arose as SP SINEs in a common ancestor of Supraprimates, comprising the mammalian orders of Primates, Dermoptera, Scandentia (grouped as Eurarchonta) as well as Lagomorpha and Rodentia (grouped as Glires) [57]. In primates as in most other orders of Supraprimates, monomeric Alu RNAs were the initial template(s) for Alu

SINEs before monomers had been superseded by dimeric Alu RNAs serving as templates for the highly abundant dimeric Alu SINEs [58, 59]. The first master gene of dimeric Alu elements came about by fusion of an SP SINE to a FAM/FRAM-derived sequence that arose, presumably independently, in the lineage leading to primates [57].

Like all SINEs, the vast majority was transcriptionally inactive after retroposition. A rare transcriptionally active monomer led to BC200 RNA. Importantly, expression persisted in all Anthropoidea lineages (New World monkeys, Old World monkeys and Apes) for 35–55 million years. Interestingly, BC200 RNA is the parent of several hundred BC200-derived SINEs or pseudogenes [60]. The genesis of Alu SINEs and BC200 RNA is depicted in figure 2. The gene encoding BC200 RNA is located on human chromosome 2 to band 2p21 between the *CALM2* (calmodulin 2) and the *TACSTD1* (tumour-associated calcium signal transducer 1) genes and is absent in prosimians (but see below). Like its rodent counterpart, BC200 RNA (~200 nt) has a tripartite structure whereby the 5′ domain (~120 nt) and the central adenosine-rich domain originated from the Alu SINE and the 3′ domain from the locus of integration and is expressed in neurons and transported to dendrites [61]. Also like BC1 RNA, low levels of expression also occur in testes [56, 60], and deregulation in immortalised cell cultures as well as in certain tumours is observed [62]. The 5′ domain BC200 folds into a secondary structure similar to SRP RNA [56] and hence it is not surprising that the protein dimer SRP9/14 binds to BC200 RNA in vitro and in vivo [63, 64]. Also, as expected from the presence of an adenosine-rich region, BC200 RNA binds in vitro and in vivo to PABP [51]. Likewise, the adenosine-rich region is responsible for inhibition of translation when tested in rabbit reticulocyte lysate [52].

The BC200 gene is located between two Alu elements of the same subfamily (Alu_{sx}). Segmental deletions often occur between sequences that are highly similar, including Alu elements. We searched in genomic DNA from 600 male patients with reproductive deficiencies. None had a detectable deletion in the BC200 RNA locus.

In phylogenetic studies, not only the BC200 RNA loci of Anthropoidea were sequenced but also those of three prosimian species. Apart from a representative of Lemuriformes and Lorisiformes (Strepsirhini) each, a representative of Tarsoidea was sequenced; the latter branched off prior to New World monkeys on the lineage leading to humans [60]. The tarsier locus turned out to be devoid of the gene encoding BC200 RNA.

Surprisingly, in Strepsirhini the locus revealed a related yet different SINE integration, namely a dimeric Alu element [60, 65]. It was ruled out that the monomeric BC200 RNA gene arose from deletion of a dimeric Alu half in Anthropoidea and its absence in tarsier is due to precise excision of a primordial

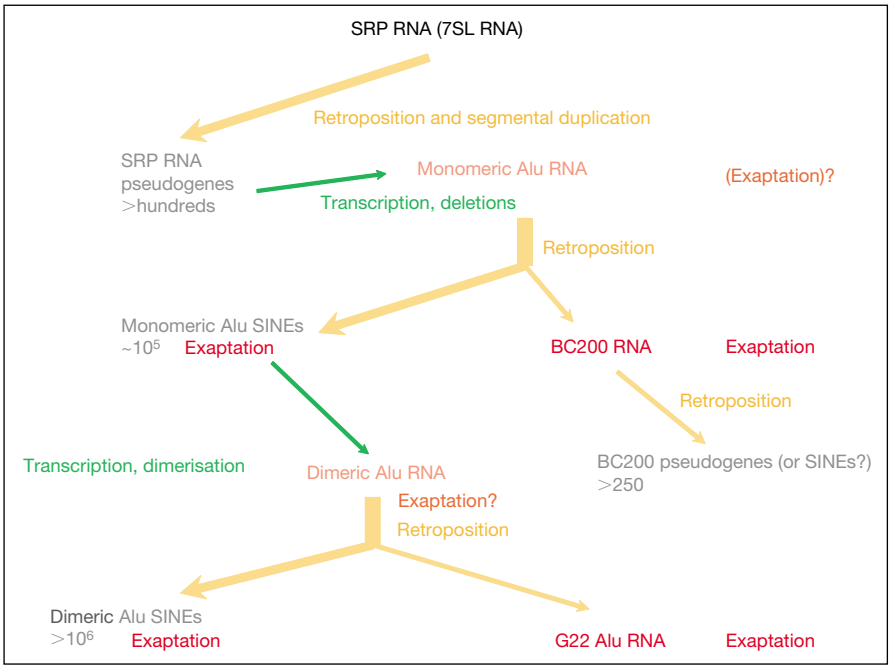


Fig. 2. Biogenesis of Alu elements, BC200 RNA and derived retrosequences. A master gene for monomeric Alu elements had probably been generated from SRP RNA by retroposition or segmental gene duplication in a common ancestor of Supraprimates. Apart from BC200 RNA (see below), it is not clear whether the corresponding monomeric Alu RNA had ever been exapted into a function but, possibly, several master genes were active over time. In any event, subsequently a transcribed master gene for dimeric Alu elements had been generated by fusion of two different Alu monomers. Again, over time, several master genes were active. Whether dimeric Alu RNAs other than G22 Alu RNA from Lorisoidea ever were exapted, is not clear. In any event, numerous Alu elements (monomeric and dimeric) have been exapted into coding or regulatory functions in the genome. In addition, one monomeric Alu element was exapted as neural BC200 npcRNA. This RNA served as template to several hundred retropseudogenes (or SINEs).

dimeric Alu element [65]. This constitutes an independent SINE integration into precisely the same locus. Northern blot analysis has shown that the dimeric Alu RNA (G22 Alu RNA) is not transcribed in *Lemur coronatus* brain but in brain of *Galago moholi* [65]. This is consistent with conservation of CpG dimers in *Galago* but not in *Lemur*. Transcriptional studies by transfecting the loci of five additional Lemuriform species into HeLa cells showed absence of

activity while three additional Lorisiform species showed transcription of G22 Alu RNA [65]. Transgenic mice express *Galago moholi* G22 Alu RNA (provided that sufficient flanking regions are present) and the RNA is found in dendrites, analogous to human BC200 RNA transgenes [66].

Most likely, a dimeric Alu element inserted into the empty BC200/G22 locus in the common ancestor of Strepsirhini after divergence of the Haplorhini branch. Either the locus was immediately transcribed yielding G22 Alu RNA, but this activity did not persist in Lemuriformes, or transcription was activated later only in Lorisiformes. A more likely scenario is the former, because the independent retroposition event of a monomeric Alu element in Anthropoidea apparently generated an active RNA-coding gene, BC200 RNA. Two separate events that initially did not transcribe the Alu elements would have required parallel and similar changes for transcriptional activation in both lineages.

Retroposed Copies of snoRNAs and Pre-miRNAs

Expression of an Alu-related RNA in the brain and its transport in dendrites must confer a selective advantage to Anthropoidea and Lorisoidea but not (any more) to Lemuroidea. Our examples show that the chances for an 'active life' of RNA polymerase III transcribed retrogenes might be slim, as not only flanking sequences have to be acquired but they also have to be located at the right distance to internal promoter elements, such as box A and box B in Alu elements. The paucity of independent Alu transcripts despite the potential of $>10^6$ copies in the genome underscores these requirements. One would predict then, that RNAs transcribed by RNA polymerase II primary transcripts with subsequent processing, if retroposed, would have a higher chance to an 'active life' as, often, they could integrate into introns and be co-transcribed with primary transcripts in the novel location and processed from introns as is the case with small nuclear RNAs (snoRNA) [67–69].

Gene duplication of snoRNA by retroposition was suggested a few years ago [70] and bioinformatic evidence is beginning to accumulate [71–73]. Amplification by retroposition could be especially feasible for RNAs that are being processed exonucleolytically from introns [74]. This way, the mature RNA can, perhaps after atypical polyadenylation, serve as template for retroposition. If integrated into an intron, generation of a functional snoRNA copy is likely (fig. 3). One could also imagine that partially processed snoRNAs will serve as templates for retroposition. Variant and (after sufficient time for changes) 'novel' micro RNAs (miRNAs) also keep arising in various lineages [75, 76]. In addition to segmental duplication as a mechanism of amplification

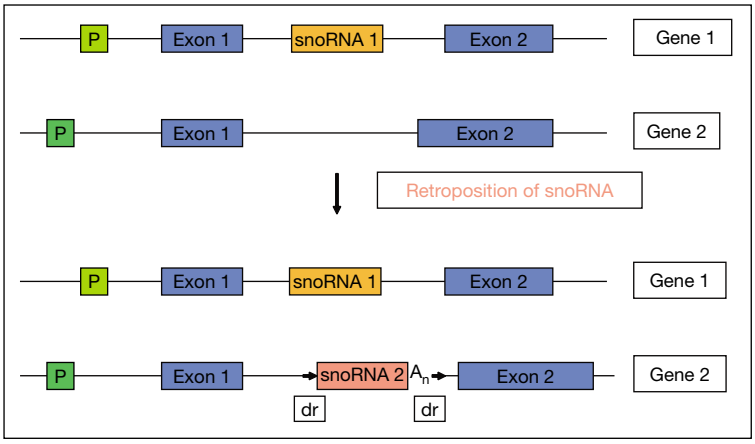


Fig. 3. Generation of novel snoRNA genes by retroposition. The two lines at the top depict two different genes on two separate chromosomal loci. Promoters (P) are depicted in green, exons in blue and introns by lines. Gene 1 hosts a snoRNA gene (orange). Processing of the hnRNA primary transcript not only generates mRNA but also snoRNA 1. snoRNA 1 is being retroposed and fortuitously integrates into an intron of the second gene. If the host gene 2 is expressed in different cell types and/or at different times in development as the gene that hosts snoRNA 1, a consequence is that the snoRNA 2 isoform is expressed like its new host. The snoRNA 2 retrogene leaves two hallmarks of retroposition, short direct flanking repeats and an adenosine-rich region at the 3′ end, presumably arisen from atypical polyadenylation prior to retroposition. Over time, such hallmarks disappear by base changes – likewise snoRNA 2 will differ more and more from its founder, snoRNA 1, possibly even changing complementarity towards a different target.

one could also envision retroposition. However, this would not involve the 21–23 nt long mature miRNAs but rather the hairpin-structured precursor miRNAs (pre-miRNA) or even longer parts of the miRNA primary transcripts (fig. 4). Searches for retroposon hallmarks around genes encoding lineage-specific pre-miRNAs should reveal cases that had been amplified via RNA intermediates.

The Tip of the Iceberg

Not even two decades ago, any nucleotide sequence that was generated by retroposition, whether retropseudogenes, SINEs or LINEs, were considered genomic waste material. Alone the contribution of retrogenes to novel protein genes (or parts thereof) is remarkable. It should be emphasized once more that

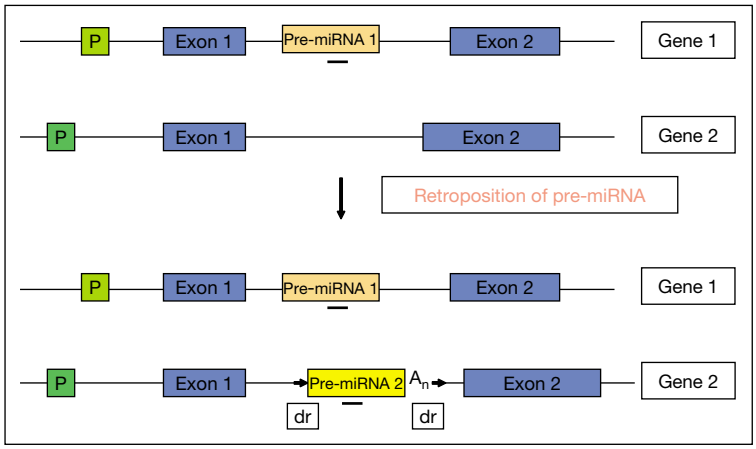


Fig. 4. Generation of novel miRNA precursor genes by retroposition. The two lines at the top depict two different genes on two separate chromosomal loci. Promoters (P) are depicted in green, exons in blue and introns by lines. Gene 1 hosts a pre-miRNA gene (ochre). Processing of the hnRNA primary transcript not only generates mRNA but also pre-miRNA 1 and eventually miRNA 1. Pre-miRNA 1 is being retroposed and fortuitously integrates into an intron of the second gene. If the host gene 2 is expressed in different cell types and/or at different times in development as the gene that hosts miRNA 1, a consequence is that the miRNA 2 isoform is expressed like its new host. The pre-miRNA 2 retrogene (yellow) is predicted to leave two hallmarks of retroposition, short direct flanking repeats and an adenosine-rich region at the 3′ end, presumably arisen from atypical polyadenylation prior to retroposition. Over time, such hallmarks disappear by base changes – likewise pre-miRNA 2 will differ more and more from its founder, pre-miRNA 1, possibly even changing complementarity towards a different mRNA target.

any type of gene duplication – if it results in an active gene – initially provides a second, usually identical copy, albeit the retrogene is likely to be expressed in different cell types and/or at different times in development. With time, one copy changes to yield a gene product that is an isoform or variant. Over longer evolutionary periods, one of the copies might acquire so many changes that its relation to the parent gene is not discernible any more. This is when the old retrogene becomes a ‘novel’ gene. In a similar vein, many genes or parts of genes, whether protein-coding or non-protein-coding, that are derived from retrotransposed elements, are not discernible as such any more, because they lost their hallmarks over time. Hence, the events that are identifiable today are the mere tip of the iceberg. However, these events are sufficient to demonstrate how retroposition contributed in shaping the genic landscapes mainly of genomes from multicellular organisms.

Acknowledgements

We apologise to the colleagues whose work could not be cited due to space limitation. J.N.V. was supported by the Biofuture program of the Bundesministerium für Bildung und Forschung (BMBF), the Association pour la Recherche contre le Cancer (ARC) and the Centre National de la Recherche Scientifique (CNRS), J.B. by grants from European Union (EU; LSHG-CT-2003-503022) and Nationales Genomforschungsnetz (NGFN 0313358A).

References

1 Orgel LE, Crick FH: Selfish DNA: the ultimate parasite. Nature 1980;284:604–607.
2 Brosius J: RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 1999;238:115–134.
3 Kazazian HH Jr: Mobile elements: drivers of genome evolution. Science 2004;303:1626–1632.
4 Nekrutenko A, Li WH: Transposable elements are found in a large number of human protein-coding genes. Trends Genet 2001;17:619–621.
5 Gotea V, Makalowski W: Do transposable elements really contribute to proteomes? Trends Genet 2006;22:260–267.
6 Krull M, Brosius J, Schmitz J: Alu-SINE exonization: en route to protein-coding function. Mol Biol Evol 2005;22:1702–1711.
7 Marino-Ramirez L, Lewis KC, Landsman D, Jordan IK: Transposable elements donate lineage-specific regulatory sequences to host genomes. Cytogenet Genome Res 2005;110:333–341.
8 Debarry JD, Ganko EW, McCarthy EM, McDonald JF: The contribution of LTR retrotransposon sequences to gene evolution in *Mus musculus*. Mol Biol Evol 2006;23:479–481.
9 Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr: Mobile elements and mammalian genome evolution. Curr Opin Genet Dev 2003;13:651–658.
10 Volff JN: Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 2006;28:913–922.
11 Brandt J, Schrauth S, Veith AM, Froschauer A, Haneke T, et al: Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. Gene 2004;345:101–111.
12 Youngson NA, Kocialkowski S, Peel N, Ferguson-Smith AC: A small family of sushi-class retrotransposon-derived genes in mammals and their relation to genomic imprinting. J Mol Evol 2005;61:481–490.
13 Poulter R, Butler M: A retrotransposon family from the pufferfish (fugu) *Fugu rubripes*. Gene 1998;215:241–249.
14 Wills NM, Moore B, Hammer A, Gesteland RF, Atkins JF: A functional -1 ribosomal frameshift signal in the human paraneoplastic *Ma3* gene. J Biol Chem 2006;281:7082–7088.
15 Nagasaki K, Schem C, von Kaisenberg C, Biallek M, Rosel F, et al: Leucine-zipper protein, LDOC1, inhibits NF-kappaB activation and sensitizes pancreatic cancer cells to apoptosis. Int J Cancer 2003;105:454–458.
16 Tan KO, Tan KM, Chan SL, Yee KS, Bevort M, et al: MAP-1, a novel proapoptotic protein containing a BH3-like motif that associates with Bax through its Bcl-2 homology domains. J Biol Chem 2001;276:2802–2807.
17 Ono R, Nakamura K, Inoue K, Naruse M, Usami T, et al: Deletion of *Peg10*, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nat Genet 2006;38:101–106.
18 Seitz H, Youngson N, Lin SP, Dalbert S, Paulsen M, et al: Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. Nat Genet 2003;34:261–262.
19 Best S, Le Tissier P, Towers G, Stoye JP: Positional cloning of the mouse retrovirus restriction gene *Fv1*. Nature 1996;382:826–829.
20 Campillos M, Doerks T, Shah PK, Bork P: Computational characterization of multiple Gag-like human proteins. Trends Genet 2006;22:585–589.

21 Llorens C, Marin I: A mammalian gene evolved from the integrase domain of an LTR retrotransposon. Mol Biol Evol 2001;18:1597–1600.
22 Dlakic M: A model of the replication fork blocking protein Fob1p based on the catalytic core domain of retroviral integrases. Protein Sci 2002;11:1274–1277.
23 Mi S, Lee X, Li X, Veldman GM, Finnerty H, et al: Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 2000;403:785–789.
24 Blaise S, de Parseval N, Benit L, Heidmann T: Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies *syncytin 2*, a gene conserved on primate evolution. Proc Natl Acad Sci USA 2003;100:13013–13018.
25 Dupressoir A, Marceau G, Vernochet C, Benit L, Kanellopoulos C, et al: Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci USA 2005;102:725–730.
26 Dunlap KA, Palmarini M, Varela M, Burghardt RC, Hayashi K, et al: Endogenous retroviruses regulate periimplantation placental growth and differentiation. Proc Natl Acad Sci USA 2006;103:14390–14395.
27 Malik HS, Henikoff S: Positive selection of *Iris*, a retroviral envelope-derived host gene in *Drosophila melanogaster*. PLoS Genet 2005;1:e44.
28 Nakamura TM, Cech TR: Reversing time: origin of telomerase. Cell 1998;92:587–590.
29 Eickbush TH: Telomerase and retrotransposons: which came first? Science 1997;277:911–912.
30 Brosius J: Many G-protein-coupled receptors are encoded by retrogenes. Trends Genet 1999;15:304–305.
31 Vinckenbosch N, Dupanloup I, Kaessmann H: Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA 2006;103:3220–3225.
32 Brosius J: Retroposons – seeds of evolution. Science 1991;251:753.
33 Betran E, Thornton K, Long M: Retroposed new genes out of the X in *Drosophila*. Genome Res 2002;12:1854–1859.
34 Emerson JJ, Kaessmann H, Betran E, Long M: Extensive gene traffic on the mammalian X chromosome. Science 2004;303:537–540.
35 Bradley J, Baltus A, Skaletsky H, Royce-Tolland M, Dewar K, Page DC: An X-to-autosome retrogene is required for spermatogenesis in mice. Nat Genet 2004;36:872–876.
36 Brosius J, Gould SJ: On ‘genomenclature’: a comprehensive (and respectful) taxonomy for pseudogenes and other ‘junk DNA’. Proc Natl Acad Sci USA 1992;89:10706–10710.
37 Brosius J: Echoes from the past – are we still in an RNP world? Cytogenet Genome Res 2005;110:8–24.
38 Brosius J: More *Haemophilus* and *Mycoplasma* genes. Science 1996;271:1302.
39 Ambros V: microRNAs: tiny regulators with great potential. Cell 2001;107:823–826.
40 Mattick JS, Makunin IV: Non-coding RNA. Hum Mol Genet 2006;15(spec 1):R17–R29.
41 Sutcliffe JG, Milner RJ, Gottesfeld JM, Lerner RA: Identifier sequences are transcribed specifically in brain. Nature 1984;308:237–241.
42 DeChiara TM, Brosius J: Neural BC1 RNA: cDNA clones reveal nonrepetitive sequence content. Proc Natl Acad Sci USA 1987;84:2624–2628.
43 Kim J, Martignetti JA, Shen MR, Brosius J, Deininger P: Rodent BC1 RNA gene as a master gene for ID element amplification. Proc Natl Acad Sci USA 1994;91:3607–3611.
44 Kim J, Deininger PL: Recent amplification of rat ID sequences. J Mol Biol 1996;261:322–327.
45 Martignetti JA, Brosius J: BC1 RNA: transcriptional analysis of a neural cell-specific RNA polymerase III transcript. Mol Cell Biol 1995;15:1642–1650.
46 Tiedge H, Fremeau RT Jr, Weinstock PH, Arancio O, Brosius J: Dendritic location of neural BC1 RNA. Proc Natl Acad Sci USA 1991;88:2093–2097.
47 Rozhdestvensky TS, Kopylov AM, Brosius J, Huttenhofer A: Neuronal BC1 RNA structure: evolutionary conversion of a tRNA(Ala) domain into an extended stem-loop structure. RNA 2001;7:722–730.
48 Muslimov IA, Iacoangeli A, Brosius J, Tiedge H: Spatial codes in dendritic BC1 RNA. J Cell Biol 2006;175:427–439.
49 Muslimov IA, Banker G, Brosius J, Tiedge H: Activity-dependent regulation of dendritic BC1 RNA in hippocampal neurons in culture. J Cell Biol 1998;141:1601–1611.

50 Chen W, Heierhorst J, Brosius J, Tiedge H: Expression of neural BC1 RNA: induction in murine tumours. Eur J Cancer 1997;33:288–292.
51 Muddashetty R, Khanam T, Kondrashov A, Bundman M, Iacoangeli A, et al: Poly(A)-binding protein is associated with neuronal BC1 and BC200 ribonucleoprotein particles. J Mol Biol 2002;321: 433–445.
52 Kondrashov AV, Kiefmann M, Ebnet K, Khanam T, Muddashetty RS, Brosius J: Inhibitory effect of naked neural BC1 RNA or BC200 RNA on eukaryotic in vitro translation systems is reversed by poly(A)-binding protein (PABP). J Mol Biol 2005;353:88–103.
53 Kindler S, Wang H, Richter D, Tiedge H: RNA transport and local control of translation. Annu Rev Cell Dev Biol 2005;21:223–245.
54 Lewejohann L, Skryabin BV, Sachser N, Prehn C, Heiduschka P, et al: Role of a neuronal small non-messenger RNA: behavioural alterations in BC1 RNA-deleted mice. Behav Brain Res 2004;154:273–289.
55 Watson JB, Sutcliffe JG: Primate brain-specific cytoplasmic transcript of the Alu repeat family. Mol Cell Biol 1987;7:3324–3327.
56 Skryabin BV, Kremerskothen J, Vassilacopoulou D, Disotell TR, Kapitonov VV, et al: The BC200 RNA gene and its neural expression are conserved in Anthropoidea (Primates). J Mol Evol 1998;47:677–685.
57 Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J: Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. Trends Genet 2007;23:153–161.
58 Shen MR, Batzer MA, Deininger PL: Evolution of the master Alu gene(s). J Mol Evol 1991;33: 311–320.
59 Jurka J, Milosavljevic A: Reconstruction and analysis of human Alu genes. J Mol Evol 1991;32: 105–121.
60 Kuryshev VY, Skryabin BV, Kremerskothen J, Jurka J, Brosius J: Birth of a gene: locus of neuronal BC200 snmRNA in three prosimians and human BC200 pseudogenes as archives of change in the Anthropoidea lineage. J Mol Biol 2001;309:1049–1066.
61 Tiedge H, Chen W, Brosius J: Primary structure, neural-specific expression, and dendritic location of human BC200 RNA. J Neurosci 1993;13:2382–2390.
62 Chen W, Bocker W, Brosius J, Tiedge H: Expression of neural BC200 RNA in human tumours. J Pathol 1997;183:345–351.
63 Bovia F, Wolff N, Ryser S, Strub K: The SRP9/14 subunit of the human signal recognition particle binds to a variety of Alu-like RNAs and with higher affinity than its mouse homolog. Nucleic Acids Res 1997;25:318–326.
64 Kremerskothen J, Zopf D, Walter P, Cheng JG, Nettermann M, et al: Heterodimer SRP9/14 is an integral part of the neural BC200 RNP in primate brain. Neurosci Lett 1998;245:123–126.
65 Ludwig A, Rozhdestvensky TS, Kuryshev VY, Schmitz J, Brosius J: An unusual primate locus that attracted two independent Alu insertions and facilitates their transcription. J Mol Biol 2005;350: 200–214.
66 Khanam T, Rozhdestvensky TS, Bundman M, Galiveti CR, Handel S, et al: Two primate-specific small non-protein-coding RNAs in transgenic mice: neuronal expression, subcellular localization and binding partners. Nucleic Acids Res 2007;35:529–539.
67 Leverette RD, Andrews MT, Maxwell ES: Mouse U14 snRNA is a processed intron of the cognate hsc70 heat shock pre-messenger RNA. Cell 1992;71:1215–1221.
68 Tycowski KT, Shu MD, Steitz JA: A small nucleolar RNA is processed from an intron of the human gene encoding ribosomal protein S3. Genes Dev 1993;7:1176–1190.
69 Balakin AG, Smith L, Fournier MJ: The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions. Cell 1996;86:823–834.
70 Brosius J: The contribution of RNAs and retroposition to evolutionary novelties. Genetica 2003;118:99–116.
71 Vitali P, Royo H, Seitz H, Bachellerie JP, Huttenhofer A, Cavaille J: Identification of 13 novel human modification guide RNAs. Nucleic Acids Res 2003;31:6543–6551.
72 Luo Y, Li S: Genome-wide analyses of retrogenes derived from the human box H/ACA snoRNAs. Nucleic Acids Res 2007;35:559–571.

73 Weber MJ: Mammalian small nucleolar RNAs are mobile genetic elements. PLoS Genet 2006;2:e205.
74 Kiss T, Filipowicz W: Exonucleolytic processing of small nucleolar RNAs from pre-mRNA introns. Genes Dev 1995;9:1411–1424.
75 Houbaviy HB, Dennis L, Jaenisch R, Sharp PA: Characterization of a highly variable eutherian microRNA gene. RNA 2005;11:1245–1257.
76 Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, et al: Diversity of microRNAs in human and chimpanzee brain. Nat Genet 2006;38:1375–1377.

Jean-Nicolas Volff
Equipe 'Génomique Evolutive des Vertébrés'
Institut de Génomique Fonctionnelle de Lyon
UMR5242 CNRS/INRA/Université Claude Bernard
LyonI/ENS Ecole Normale Supérieure de Lyon
46 allée d'Italie, F-69364 Lyon Cedex 07 (France)
Tel. +33 4 72 72 81 16, Fax +33 4 72 72 86 99, E-mail Jean-Nicolas.Volff@ens-lyon.fr

Author Index

Albà, M.M. 119
Aravind, L. 48, 66
Arguello, J.R. 131

Balaji, S. 66
Blomme, T. 81
Bowen, N.J. 147
Brosius, J. 175
Burroughs, A.M. 48

Dutheil, J. 1

Fan, C. 131

Galtier, N. 1
Gojobori, T. 13
Gophna, U. 30

Iyer, L.M. 48

Jordan, I.K. 147

Long, M. 131

Madan Babu, M. 66
Makałowski, W. 163
Makino, T. 13
Marchal, K. 81

Pallen, M.J. 30
Petrov, D.A. 101

Singh, N.D. 101

Toda, Y. 163
Tompa, P. 119

Van de Peer, Y. 81
Van Hellemont, R. 81
Veitia, R.A. 119
Volff, J.-N. 175

Wang, W. 131

Subject Index

Alu sequence 141, 149, 165
Amino acid repeats 119
Amyloid 125
Antisense RNA 137
ATPase 31, 50

Background selection 102
Bacterial flagella 30
BC1 RNA 179
BC200 RNA 181
Biological diversity 131
Biological network 13
Bone morphogenetic protein (Bmp) 91

Calmodulin 22
Capsid 49
Chimeric genes 131
Chimeric protein 135
Chromatin-remodeling complex 21
Chromosomal heteromorphy 105
Coding repeat 120
Coevolution 1, 69
Comparative genomics 48, 131
Complexity 33
Condon bias 109
Correlated patterns 2
Correlated processes 4

Darwinian evolution 30
Dense part protein 24
Different function protein 23
DNA packaging system 48
DNA transposon 149, 166
Dobzhansky-Muller incompatibility 7
Double strand break 133

Effective population size 109
Envelope protein 177
Epistasis 2
Evolutionary distance 25
Evolutionary novelty 131
Evolutionary rate 16, 105
Exaptation 147, 167
Exonisation 168, 176
Exon shuffling 133

Faster-X hypothesis 106
Feed-forward motif 71
Female-biased genes 111
Fish-specific genome duplication 81
Flagellar motor 42
Flagellin 35

Gag protein 177
GC content 122
Gene dosage 104
Gene duplication 5, 15, 34, 81, 111, 132, 169, 184
Gene fusion 135
Gene traffic 111
Genetic hitchhiking 103
Global network structure 74
Global transcription activator complex 21

Hedgehog genes 94
Helitron 133

Helix-turn-helix domain 57, 157
Homologous recombination 133
Homopeptide 120
Homopolymeric run 119
Human disease 120
Human tissue plasminogen activator 136

Imprinting 177
Intelligent design 32

Junk DNA 163

Local network structure 71
Long interspersed element (LINE) 141, 165, 176
Long terminal repeat retrotransposon 148, 164, 176
Low copy repeat 133
Low density lipoprotein receptor 136

Male-biased genes 111
MicroRNA 171, 184
Microsatellite 120
Miniature inverted repeat element (MITE) 149
Mirror tree method 5
Modularity 33, 170
Molecular bricolage 34
Molecular domestication 151
Muller's ratchet 102
Multiple-input motif 71

Nearly perfect repeats 124
Neofunctionalization 81
Network motif 71
Non-homologous recombination 133
Non-allelic homologous recombination 133

Operon 56

Pack-MULE 133
Parasite 70
Paralogous intergenic region 89
Pathogen 75
Pax6 92
Phage 50
Phylogenetic footprinting 88
Phylogeny 3, 36, 59

P-loop NTPase fold 52
Polyalanines 119
PolyA-binding protein 181
Polyglutamines 119
Portal protein 49
Positive selection 109
Promoter 170
Protein-protein interaction 13, 30
Protein structure 119

RAG recombinase 153
Recombination 35, 102, 120, 132, 147, 153
Regulatory hub 74
Regulatory interaction 69
Regulatory network 66
Regulatory protein 67
Regulatory sequence 81
Relative simplicity factor 122
Repetition 33
Replication slippage 120
Retroelement 148, 175
Retrogene 112, 132, 166, 178
Retro(trans)position 111, 132, 175, 184
Ribonucleoprotein 181
Ribosomal RNA 7
RNA gene 179
RNA/RNP world 179

Same function protein 23
Selfish DNA theory 150
SETMAR protein 153
Sex chromosome 101
Short interspersed element (SINE) 141, 148, 165, 176
Single input motif 71
Small nuclear RNA 184
Sparse part protein 24
Spermatogenesis 111, 179
Src homolgy 3 fold 55
Subfunctionalization 81
SVA element 166
Syncytin 178

Target gene 69
Telomerase 153, 178
Terminase 49
Topology 54, 67
Transcriptional network 66

Transcription factor 67, 122, 157
Transcription factor binding site 82
Translocation 112, 140
Transposable element 133, 147, 163, 176
Transposable element cassette 167
Trinucleotide repeat expansion 120
Type III secretion 30

V(D)J recombination 153
Virus 49

Weak selection Hill-Roberston effect 103

X chromosome 101
X chromosome inactivation 111
X-linked genes 103, 179

Y chromosome degeneration 103